REGIANE TEODORO DO AMARAL

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
2015

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

A485n
2015

Amaral, Regiane Teodoro do, 1971-
Números de repetições na identificação de genes diferencialmente expressos em experimentos RNA-Seq / Regiane Teodoro do Amaral. - Viçosa, MG, 2015.
viii, 48f. : il. (algumas color.) ; 29 cm.

Inclui apêndice.
Orientador: Moysés Nascimento.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Estatística aplicada. 2. Biometria. 3. Biologia molecular - Métodos estatísticos. 4. Transcriptoma. 5. 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

REGIANE TEODORO DO AMARAL

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: 27 de fevereiro de 2015.

Talles Eduardo Ferreira Maciel

Ana Carolina Campana Nascimento
(Coorientadora)

Moysés Nascimento
(Orientador)

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes”. (Martin Luther King)

AGRADECIMENTOS

A Deus, pela força para ultrapassar os momentos difíceis e pela inspiração necessária para chegar ao final desta etapa.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, por proporcionar a realização de um curso de excelência.

À minha família, pelo carinho e incentivo em todos os momentos.

Ao orientador Moysés Nascimento, pelos ensinamentos, confiança, dedicação e por contribuir para o meu crescimento profissional.

Aos membros da banca examinadora, Talles Eduardo Ferreira Maciel e Ana Carolina Campana Nascimento, pela disponibilidade e pelas sugestões para o enriquecimento deste trabalho.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, por contribuírem para minha formação acadêmica.

Aos funcionários, do Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela prontidão.

À FAPEMIG, pela concessão da bolsa de estudos.

E a todos que de alguma forma contribuíram para o meu crescimento profissional e para a concretização deste trabalho.

BIOGRAFIA

REGIANE TEODORO DO AMARAL, filha de José Teodoro Sobrinho e Lourdes Luiza Teodoro, nasceu em Nova Venécia, Espírito Santo, em 24 de abril de 1971.

Em abril de 2001, ingressou no curso de Bacharelado em Estatística na Universidade Federal do Espírito Santo, Vitória – ES, graduando-se em agosto de 2006.

Em abril de 2013, iniciou o curso de Mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 27 de fevereiro de 2015.

SUMÁRIO

RESUMO

AMARAL, Regiane Teodoro do, M.Sc., Universidade Federal de Viçosa, fevereiro de 2015. **Número de repetições na identificação de genes diferencialmente expressos em experimentos de RNA-Seq.** Orientador: Moysés Nascimento. Coorientadores: Luiz Alexandre Peternelli, Ana Carolina Campana Nascimento e Fabyano Fonseca e Silva.

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. O método RNA-Seq usa transcriptoma a partir de tecnologias de sequenciamentos de nova geração (SNG), utilizados para sequenciar cDNA que é derivado de uma amostra de RNA, e, assim, produzir milhões de sequenciamentos de leitura. Porém, apesar do custo dessas tecnologias vir diminuindo, é comum realizar experimentos com pouca ou nenhuma repetição. 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 na comparação de metodologias estatísticas a fim de estudar o padrão de expressão gênica relacionado à quantificação desses genes conforme determinadas condições/tratamentos, em experimentos de RNA-Seq. Para a realização das análises utilizou-se um conjunto de dados simulados através do pacote TCC do R, com diferentes cenários, para comparar os métodos estatísticos DESeq e baySeq. Foram exploradas tecnologias de RNA-Seq do perfil de expressão gênica de um banco de dados contendo 1000 genes em duas condições, nos cenários com cinco repetições, três repetições, 2 repetições e sem repetição. Em um primeiro momento, tais dados foram analisados pelos dois métodos separadamente, comparando-se o efeito do número de repetições dentro de cada um. Em seguida, foi realizada a comparação entre os métodos, levando em conta também o número de repetições em cada cenário. De acordo com os resultados gerados nas análises não podemos afirmar que um método, entre os avaliados, é ótimo em todas as circunstâncias, pois o método de escolha para uma situação em particular depende das condições experimentais. No entanto, sob as condições utilizadas no desenvolver do experimento, o método abordado pelo baySeq foi o que apresentou um bom desempenho, nas combinações ocorridas entre os métodos e os tipos de genes analisados, ou seja, esse foi o método que obteve uma maior capacidade de identificação dos genes diferencialmente expressos.

ABSTRACT

AMARAL, Regiane Teodoro do, M.Sc., Universidade Federal de Viçosa, February, 2015. **Number of repetitions in identifying differentially expressed genes in RNA-Seq experiments.** Advisor: Moysés Nascimento. Co-Advisors: Luiz Alexandre Peternelli, Ana Carolina Campana Nascimento and Fabyano Fonseca e Silva.

One of the main challenges of molecular biology is to measure and assess the gene expression profiles in different biological tissues in order to understand the molecular mechanisms of transformation. The method uses RNA Seq transcriptome from Young generation sequencing technologies (NGS), used to sequence the cDNA which is derived from an RNA sample, and thus produce millions of reading sequencing. However, despite the cost of these technologies come decreasing, it is common experiment with little or no repetition. Thus, it becomes necessary discovery and improvement of efficient statistical methods to optimize the data analysis generated genome sequencing platforms. The aim of this study was to compare statistical methodologies to study the pattern of gene expression related to the quantification of these genes as certain conditions / treatments in RNA-Seq experiments. To carry out the analysis used a set of simulated data via the R TCC package with different scenarios to compare the statistical methods DESeq and baySeq. RNA-Seq technology of gene expression profile of a database containing 1000 genes were explored in two groups, in scenarios with five repetitions, three replicates, 2 repetitions and without repetition. At first, these data were analyzed by two methods separately, comparing the effect of the number of repetitions within each. Then, the comparison between the methods was carried out, taking into account also the number of repetitions in each scenario. According to the results generated in the analyzes can not be said that a method, among the evaluated, is great in all circumstances, as the method of choice for a particular situation depends on the experimental conditions. However, under the conditions used in developing the experiment, the method was approached by baySeq which performed well, in combinations that occurred between the methods and the types of genes analyzed, that is, that was the method that obtained a greater capacity Identification of differentially expressed genes.

1 INTRODUÇÃO GERAL

O transcriptoma é o conjunto completo de transcritos (RNAs mensageiros, RNAs ribossômicos, RNAs transportadores e os microRNAs) de um organismo, órgão, tecido ou linhagem celular. Sendo assim, ele é o reflexo direto da expressão dos genes podendo apresentar variações em determinado momento (numa dada fase do ciclo celular, por exemplo), e estado fisiológico. O estudo do transcriptoma tem, como principais objetivos: identificar tipo de transcritos, determinar a estrutura dos genes e quantificar as mudanças nos níveis de expressão gênica de cada transcrito (WANG et al, 2009).

A partir do ano de 1995, uma das principais técnicas utilizadas para mensurar e inferir sobre transcriptomas era a de microarranjos de DNA (microarrays). A tecnologia de microarranjos é um processo baseado em hibridização que possibilita observar a concentração de mRNA de uma amostra/condição, analisando a intensidade de sinais fluorescentes, sendo a hibridização o processo bioquímico onde duas fitas de ácido nucléico, com sequências complementares, se combinam/emparelham (DANTAS, 2004).

Como alternativa aos microarranjos, surgiram modernos métodos de sequenciamento em larga escala, os quais permitem sequenciar o transcriptoma em uma única corrida (RNA-Seq). De modo geral, as metodologias de RNA-Seq se caracterizam pela conversão de RNA em uma biblioteca de fragmentos de cDNA, de forma que cada molécula pode ser sequenciada por um sequenciador de nova geração (plataformas) gerando pequenas sequências (reads) com tamanho variando entre 21 e 500 bp (WANG et al., 2009). Assim, após a normalização, que é a padronização dos dados obtidos, quanto maior a contagem de reads correspondentes a um determinado gene, em certa condição experimental (tratamento), maior a sua expressão. Deste modo, técnicas como o sequenciamento de RNA mensageiro (RNAm) em plataformas de sequenciamento de nova geração (RNA-Seq) se tornaram uma ferramenta importante, auxiliando a busca de genes responsáveis por caracteres de interesse. O RNA-Seq permite obter o perfil do transcriptoma, além de ser ponto de partida para a identificação de transcritos novos e/ou raros e é uma ferramenta poderosa para estudos de expressão gênica.

As vantagens dos estudos de RNA-Seq em relação aos de microarranjos são que a detecção dos transcritos não fica restrita somente àqueles correspondentes a uma sequência genômica pré-existente (tal como ocorre com as abordagens baseadas em hibridização), tornando o método atrativo para a pesquisa em organismos cujos genomas ainda não foram determinados. Além disso, esse tipo de estudo possibilita a mensuração da abundância dos

transcritos diretamente por meio da contagem das sequências em vez de analisar a intensidade da hibridização e o ruído de fundo (falta de contribuição de sinal devido a moléculas que não se anelam com nenhuma molécula fluorescente) é muito menor se comparado aos microarranjos de DNA. Outra vantagem, mencionada em Auer e Doerge (2010), se refere a maior simplicidade dos delineamentos experimentais quando comparados aos utilizados em microarrays.

De modo geral, nota-se que os desafios e soluções que envolviam a técnica de microarray são semelhantemente debatidos com o RNA-Seq. Por exemplo, os primeiros estudos de microarrays poucas vezes utilizavam repetições, e eles determinavam a expressão diferencial usando estatísticas simples, mas ao longo do tempo, se tornou evidente que é essencial considerar também a variabilidade entre as amostras. Assim, as experiências de microarray hoje fazem uso de procedimentos de testes estatísticos avançados, tais como os baseados em testes-t modificados (SMYTH, 2003). Da mesma forma, os estudos iniciais de RNA-Seq frequentemente utilizavam uma única fonte de RNA e a distribuição das contagens era considerada como uma Poisson (MARIONI, 2008). No entanto, quando os dados com repetições biológicas tornaram-se mais disponíveis, logo se percebeu que a variabilidade entre as medições repetidas é muitas vezes maior do que o esperado pela distribuição de Poisson, fenômeno conhecido como superdispersão (HINDE e DEMÉTRIO, 1998). Consequentemente, os métodos para lidar com a variabilidade biológica e superdispersão foram introduzidos, tais como métodos baseados em modelo binomial negativo e beta binomial negativo.

A literatura apresenta diversas metodologias para análise de expressão gênica para dados de RNA-Seq. Tais metodologias diferem quanto aos conceitos adotados e aos princípios estatísticos empregados. Como exemplo, citam-se as metodologias edgeR (ROBINSON et al., 2010), DESeq (ANDERS e HUBER, 2010), DEGSeq (WANG et al, 2010) e baySeq (HARDCASTLE e KELLY, 2010).

Devido a curta história de RNA-Seq e seu desenvolvimento contínuo, não há métodos padrão ainda disponíveis para detectar genes diferencialmente expressos com base nesses dados. Diante disso, com quantidade e diferenças entre os métodos, alguns trabalhos têm sido desenvolvidos a fim de compará-los em relação à capacidade de detectar genes diferencialmente expressos. Dentre eles podem ser citados, Kvam et al. (2012) que compararam os métodos estatísticos EdgeR, DESeq, baySeq, e o método com um modelo de Poisson de dois estágios (MPTA), em termos da sua capacidade para discriminar entre genes

diferencialmente expressos e genes não diferencialmente expressos, através de uma variedade de simulações que foram baseadas em diferentes modelos de distribuição ou de dados reais. Robles et al.(2012) compararam os métodos edgeR, DESeq e NBPSeq, avaliando o impacto do aumento da cobertura de sequenciamento para detectar genes diferencialmente expressos, contrastando com este as vantagens de se aumentar o tamanho da amostra. Além disso, também foi verificado nesse trabalho o controle da taxa de falsos positivos desses métodos. Em ambos os trabalhos, os resultados indicaram que o método baySeq tem menores taxas de falsos positivos. Verificou-se ainda que o MPTA não executa tão bem como os outros métodos quando o tamanho de amostra é pequeno.

Soneson e Delorenzi (2013) compararam onze métodos para análise de expressão diferencial de dados de RNA-Seq baseando-se nas características qualitativas dos dados normalizados e no impacto que o método de normalização causa nos resultados da análise de expressão diferencial. Ainda nesse trabalho foi investigada a influência do método de normalização na taxa de falsos-positivos (erro do tipo I) e o poder da análise de expressão diferencial. Como resultado, verificou-se que amostras muito pequenas, que ainda são comuns em experimentos de RNA-Seq, são ainda um problema para todos os métodos avaliados e quaisquer resultados obtidos em tais condições devem ser interpretados com cautela. Para tamanhos maiores de amostra, os métodos que utilizam o método de transformação limma (SMYTH et al., 2003) para estabilização da variância tiveram um melhor desempenho sob diversas condições, como faz o método SAMseq não paramétrico (LI e TIBSHIRANI, 2013). Com o estudo foi possível recomendar o método de normalização apropriado a ser utilizado na análise de expressão diferencial, não levando em conta o número de repetições.

Os trabalhos citados anteriormente não avaliaram de maneira específica os métodos que possibilitam a análises sem repetições (DESeq e Bayseq). Além disso, não avaliaram se os genes encontrados pelas diferentes metodologias são os mesmos.

Neste trabalho serão abordados os pacotes DESeq e o baySeq implementados no software livre R versão 2.15.1 (R Development Core Team, 2012), disponível em http://www.R-project.org. Tais pacotes foram escolhidos por permitirem a análise de dados provenientes de experimentos sem repetição e se baseiam na distribuição binomial negativa, porém com dois enfoques distintos, frequentista e bayesiano.

1.1 OBJETIVOS

Geral:

O objetivo geral deste trabalho é analisar o efeito do número de repetições na identificação de genes diferencialmente expressos em experimentos de RNA-Seq, comparando-se os métodos DEseq e Bayseq.

Os objetivos específicos são:

1. Fornecer um material didático a respeito das técnicas abordadas.
2. Apresentar os métodos para análise de expressão diferencial em um nível de exposição apropriado para alunos e pesquisadores interessados neste assunto com conhecimentos básicos em Probabilidade e Estatística e em Genética;
3. Apresentar os fundamentos teóricos das metodologias.
4. Indicar métodos para diferentes situações quanto ao número de repetições.
5. Verificar a concordância entre os genes diferencialmente expressos detectados em cenários diferentes.

2 REVISÃO DE LITERATURA

2.1 Sequenciamento de RNA-Seq a partir das tecnologias de sequenciamento de nova geração.

A técnica de RNA-Seq é o sequenciamento em larga escala de cDNA (DNA complementar) utilizando sequenciadores de nova geração, auxiliando na representação dos resultados dos transcriptomas analisados, gerando informações que são analisadas por softwares específicos, tornando estas informações mais claras aos pesquisadores que as utilizam em novas pesquisas e comparações de organismos.

O Sequenciamento de Nova Geração surgiu 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 possibilitam ainda, além do estudo do transcriptoma (RNA-Seq), avaliar muitos fenômenos biológicos, incluindo polimorfismo de nucleotídeo único (SNP), eventos epigenéticos, splicing alternativo e o estudo de interações proteína-DNA (WANG et. al., 2009 apud GONÇALVES, 2013).

.

A tecnologia de RNA-Seq é recente e está sendo cada vez mais utilizada, fazendo com que a mesma seja inovadora em pesquisas de transcriptomas, pois além de proporcionar uma maior sensibilidade em comparação aos microarrays, não necessita de uma lista pré-definida dos genes que se deseja detectar (BULLARD et al., 2010, MARIONI et al., 2008). A princípio, qualquer transcrito que estiver sendo expresso pode ser detectado se o experimento tiver cobertura suficiente.

Como discutido em Dillies et al. (2013), embora as análises estatísticas e procedimentos de bioinformática para RNA-Seq difiram consideravelmente quando comparadas às técnicas utilizadas em microarrays, os aspectos metodológicos se assemelham. Em particular, os transcritos fragmentados (short reads) são sequenciados e não hibridizados contra um chip, e devem ser montados ou alinhados com um genoma de referência. Ademais, apesar das diferentes tecnologias de sequenciamento estarem disponíveis, em geral todas compartilham das mesmas etapas de pré-processamento e análise dos dados de RNA-Seq. De acordo com Oshlack et al. (2010) essas etapas podem ser divididas da seguinte forma: (i) os short reads são pré-processados (por exemplo, para remover adaptadores e sequências de baixa qualidade) e, então mapeados em uma sequência de referência ou em um genoma; (ii) o nível de expressão é estimado para cada entidade biológica (por exemplo, um gene); (iii) os dados são normalizados e (iv) uma análise estatística é utilizada para identificar os genes diferencialmente expressos entre as diferentes condições.

Questionamentos a respeito de todas as etapas citadas ainda estão abertos e podem ter um forte impacto sobre a análise. No estudo de Dillies et al. (2012) pode ser encontrada uma avaliação de métodos para normalização de dados provenientes de sequenciamento de nova geração, ou seja a terceira etapa. Neste trabalho, o foco está especificamente na quarta etapa, ou seja, nos aspectos ligados a análise estatística para identificação de genes que apresentem comportamento diferencial entre condições distintas.

2.2 Etapas para análises da expressão gênica utilizando RNA-Seq

2.2.1 Limpeza dos dados

Uma etapa importante antes de efetuar a análise de expressão gênica é o tratamento dos dados brutos e esse consiste na retirada de regiões que não fazem parte do genoma do organismo sequenciado, tais como: vetores, adaptadores e DNAs provenientes de possíveis contaminações. Primeiramente, visualizam-se em programas apropriados, as estatísticas das

reads provenientes do sequenciamento. O programa mais utilizado para esta finalidade é fastQC. Na limpeza dos dados brutos, dentre os tipos possíveis de tratamentos, cita-se:

- ✓ Retirada de reads duplicadas em casos específicos.
- ✓ Retirada de adaptadores.
- ✓ Retirada de sequências super representadas (depende da análise).
- ✓ Clivagem de extremidades de reads com valor de phred menor que 20.
- ✓ Retirada de reads ou partes destas que contêm bases ambíguas, representada, no conjunto de dados, pela letra "N". Não existe um consenso no número de bases ambíguas para retirada da read.
- ✓ Retirada de reads com valor de phred médio menor que 20.
- ✓ Retirada das reads menores que um determinado tamanho.
- ✓ Retirada das reads maiores que um determinado tamanho.

A cobertura do sequenciamento influencia nas etapas que serão empregadas, uma vez que um tratamento estringente em um conjunto de dados com baixa cobertura pode resultar no reduzido número de reads, inviabilizando determinadas análises.

2.2.2 Mapeamento das reads

Para utilizar os dados de RNA-Seq com o objetivo de comparar a expressão gênica sob determinadas condições, é necessário transformar milhões de leituras (reads) em uma quantificação de expressão. O primeiro passo nesse processo, é o mapeamento em sequências gênicas (conhecidas como referências) das reads sequenciadas a partir dos fragmentos de interesse. Neste mapeamento, o intuito é encontrar o melhor local da referência onde cada read melhor se alinha.

Devido ao elevado conteúdo de DNA repetitivo na maioria dos genomas, cada read pode se alinhar em mais de um lugar na referência, influenciando a análise de expressão diferencial. Normalmente este mapeamento é efetuado utilizando os parâmetros default dos inúmeros programas que se destinam a este propósito. O conhecimento aprofundado do organismo que se está trabalhando é essencial para avaliar a necessidade de alteração dos parâmetros defaults destes programas, sendo os mais importantes os parâmetros cobertura (percentual da referência que foi "coberta" pela read) e a identidade (percentual de bases idênticas no alinhamento da referência com a read).

2.2.3 Normalização

Uma das etapas essenciais para o sucesso da análise de expressão diferencial é a normalização. Essa é necessária, tendo em vista que os dados provenientes de sequenciamento apresentam diferenças quanto ao número de leituras (reads) produzidas entre as diferentes condições a serem comparadas e viéses introduzidos pelos protocolos de preparação das bibliotecas (unidade experimental). Especificamente, tais problemas se referem a diferenças do tamanho da biblioteca entre amostras (cobertura do sequenciamento) e dentro das amostras, como por exemplo, o efeito do comprimento do transcrito (gene) e o conteúdo GC (DILLIES et al., 2013). Particularmente, maiores coberturas resultam em maiores quantidades para o total da amostra, o que influencia no número de reads mapeadas a cada transcrito. Além disso, Oshlack e Wakefield (2009) mostraram que o comprimento do transcrito causa alterações no ranking de genes diferencialmente expressos, que podem introduzir erros em análises posteriores. Geralmente, quanto maior o tamanho da referência, maior o número de reads mapeadas.

A partir do ano de 2010, muitos estudos na literatura propuseram diferentes metodologias para realizar a normalização de dados de RNA-Seq. Dentre os diversos métodos podem se destacar: Contagem total (Total Count - TC), Quartil Superior (Upper Quartile - UQ) (BULLARD et al., 2010), Mediana (Median - Med), a normalização implementada no pacote DESeq do Bioconductor (ANDERS e HUBER, 2010) e a normalização conhecida como Reads Per Kilobase per Million Mapped Reads (RPKM) (MORTAZAVI et al., 2008).

De acordo com Dillies et al. (2013), a fonte mais óbvia de variação entre as amostras é o tamanho da biblioteca (cobertura do sequenciamento). Logo, a forma mais simples de normalização entre amostras é dada pelo dimensionamento do número de reads por um fator específico que representa o tamanho da biblioteca. A seguir são apresentados, de forma sucinta, cinco diferentes métodos para calcular esses fatores de escala:

i) Contagem total (TC): O número de reads de cada transcrito é dividido pelo número total de reads mapeados (tamanho da biblioteca) associados com sua amostra (biblioteca) e multiplicado pela média geral das contagens em todas as amostras do conjunto de dados;

ii) Quartil superior (UQ): Semelhante ao método anterior, porém o número total de reads é substituido pelo quartil superior (0,75) dos valores de contagens que são diferentes de zero.

iii) Mediana (Med): Também semelhante aos métodos anteriores, sendo que o número total de reads é substituído pela mediana dos valores de contagens que são diferentes de zero.

iv) DESeq: Nesse método o fator de escala para uma determinada amostra é dado pela mediana dos valores de contagem de cada transcrito divido pela média geométrica das contagens de todas as amostras. Essa metodologia é baseada na suposição que a maioria dos transcritos não são diferencialmente expressos.

v) Reads Per Kilobase per Million mapped reads (RPKM): De acordo com Dillies et al. (2013) essa abordagem foi inicialmente proposta para possibilitar a comparação entre os transcritos de uma mesma amostra, uma vez que esse método visa efetuar a normalização levando em conta o tamanho da biblioteca e o comprimento do transcrito.

Embora interessante essa metodologia introduz viés na variância de cada transcrito. A expressão para o cálculo do RPKM é dada por:

$$\mathrm{RPKM} = \frac{10^{9}.\mathrm{C}}{\mathrm{N.L}}$$

em que C é o número de reads mapeados no transcrito, N é o número total de reads mapeados no experimento e L é a soma dos transcritos em pares de base.

2.3 Análise estatística da expressão gênica diferencial

Em análises de dados de RNA-Seq há um amplo interesse na identificação dos genes diferencialmente expressos, ou seja, aqueles que mudaram seus níveis de expressão em diferentes condições experimentais. Nesse tipo de análise, obtém-se uma medição discreta para cada gene, diferentemente das intensidades originadas no microarray, as quais proporcionam uma distribuição de intensidade contínua. Apesar das intensidades do microarray serem tipicamente log-transformadas e analisadas como variáveis aleatórias normalmente distribuídas, a transformação dos dados de contagem não é bem aproximada por distribuições contínuas. Dessa forma, modelos estatísticos adequados para dados de contagem são primordiais para retirar o máximo de informações a partir de dados de RNA-Seq (GONÇALVES, 2013).

Inicialmente a modelagem de dados de contagem de RNA-Seq fazia uso da distribuição de Poisson. Entretanto, considerando tal distribuição, a variabilidade biológica não é bem estimada, pois a distribuição possui um único parâmetro, o que é exclusivamente

determinado pela sua média, tornando assim a variância igual à média (LANGMEAD et al., 2010).

Apesar de útil, em estudos de RNA-Seq constatou-se que a suposição da distribuição de Poisson para avaliação da expressão diferencial na presença de repetições biológicas é demasiadamente restritiva, visto que a mesma irá predizer menores variações do que a encontrada nos dados. Portanto, testes estatísticos derivados desta pressuposição não controlam o erro tipo I (ROBINSON e SMYTH, 2008; ANDERS e HUBER, 2010).

Para tratar tal situação, denotada na literatura por overdispersion problem, foi proposto modelar os dados de RNA-Seq por meio da distribuição Binomial Negativa (ROBINSON e SMYTH, 2008). A seguir, serão apresentadas as distribuições utilizadas para modelar os dados de RNA-Seq.

2.3.1 Distribuição Binomial

Seja X o número de sucessos obtidos, na realização de n ensaios independentes de Bernoulli. Diz-se que X segue o modelo Binomial com parâmetros n e p e sua função de probabilidade é dada por:

$$P(X = x) = \binom{n}{x} p^{x}(1-p)^{n-x}, \qquad x = 0,1,\cdots,n.$$

em que n = tamanhos da amostra; p= probabilidade de sucesso e x= número total de sucessos.

A notação utilizada na representação de uma variável aleatória com distribuição Binomial é $X \sim bin(n, p)$.

Em estudos de RNA-Seq a variável aleatória é dada pelo número de reads alinhados (mapeados) a um determinado gene (transcrito). Entretanto, devido a grande quantidade de reads (tamanho da biblioteca) a probabilidade de ocorrência de um "sucesso" é muito pequena. Como na prática o número total de reads de uma biblioteca não é conhecido, torna-se necessária a utilização de outra distribuição de probabilidade para avaliar hipóteses a respeito da expressão diferencial entre diferentes condições experimentais.

Nesse caso, utiliza-se a distribuição de Poisson, tendo em vista que, para valores de probabilidade pequenos $(p \to 0)$ e valores de n grandes $(n \to \infty)$ a distribuição Binomial pode ser aproximada pela distribuição de Poisson (CASELLA e BERGER, 2010).

2.3.2 Modelo de Poisson

A distribuição de Poisson é uma distribuição de probabilidade discreta em que não se está interessado no número de sucessos obtidos em n tentativas como no caso da distribuição Binomial, mas sim, no número de sucessos ocorridos durante um intervalo contínuo (que pode ser um intervalo de tempo, espaço, área ou volume). Desta forma, não é possível determinar a probabilidade de ocorrência de um sucesso, mas sim a frequência média de sua ocorrência, que é denominada λ.

Considere X o número de sucessos obtidos e um intervalo de tempo ou espaço contínuo. Diz-se que X segue o modelo de Poisson com parâmetro λ e sua função de probabilidade é dada por:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \qquad x = 0,1,\cdots$$

em que x = quantidade de sucessos.

O parâmetro λ é conhecido como a taxa de ocorrência e a notação utilizada na representação de uma variável aleatória com distribuição de Poisson é $X \sim Po(\lambda)$.

Apesar de útil em estudos de RNA-Seq, a suposição da distribuição de Poisson para avaliação da expressão diferencial na presença de repetições biológicas é muito restritiva, pois essa estima menores variações do que as encontradas nos dados. Sendo assim, testes estatísticos que utilizam essa pressuposição não controlam o erro do tipo I (ROBINSON E SMITH, 2008; NAGALAKSHMI et al., 2008; ANDERS e HUBER, 2010).

A fim de tratar tal situação, denotada na literatura por overdispersion problem, foi proposto modelar os dados de RNA-Seq pela distribuição Binomial Negativa (ROBINSON e SMITH, 2008)

2.3.3 Distribuição Binomial Negativa

De acordo com Hinde e Demétrio, em dados de RNA-Seq a variância cresce mais rápido do que a média em um problema conhecido como super-dispersão, especialmente quando há replicação biológica entre as amostras. Esse problema ocasiona uma subestimação do erro amostral ao se adotar o modelo Poisson. Esse fenômeno ocorre quando a variância da variável resposta excede a variância nominal especificada pelo modelo adotado. As principais causas da presença da super-dispersão, segundo Hinde e Demétrio (1998), podem ser citadas como: as probabilidades de ocorrência não são constantes entre os indivíduos; as observações de ocorrência não são independentes entre os indivíduos; o processo adotado na coleta de

dados não é adequado; variáveis são omitidas na investigação do estudo ou a variabilidade do material utilizado no experimento é uma característica da ocorrência do fenômeno. Além dessas, a variabilidade em repetições biológicas, devido a heterogeneidade dentro de uma população de células, possível correlação entre as expressões de genes, devido à regulamentação, e outras variações não controladas.

A fim de contornar esse problema, a distribuição binomial negativa pode ser usada como uma alternativa para a distribuição de Poisson. Ela é especialmente útil para dados discretos em cuja amostra a variância excede a média amostral.

Considere uma sequência de ensaios de Bernoulli independentes, se definirmos X como o número de fracassos anteriores ao r-ésimo sucesso, a variável aleatória X segue uma distribuição Binomial Negativa com parâmetros r e p sendo $0 < p < 1$ e $r > 0$. Diz-se que $X \sim NB(r; p)$ e sua função de probabilidade é dada por:

$$P(X = x) = \binom{x + r - 1}{r - 1} p^r (1 - p)^x \qquad x = 0,1,\cdots$$

em que p = probabilidade de sucessos, x = quantidade de fracassos e r = número de sucessos.

Além disso, pode ser demonstrado que $E(X) = \frac{r(1-p)}{p}$ e $V(X) = \frac{r(1-p)}{p^2}$.

Outra maneira de obtenção da distribuição Binomial Negativa é com a utilização de misturas de distribuições de probabilidade. Especificamente, a distribuição Binomial Negativa é resultado da mistura das distribuições de Poisson(λ) e Gamma(α, β), onde se considera que o parâmetro λ varia de acordo com uma distribuição Gamma. Madeira et al. (2009) demonstraram que essa mistura de distribuições resulta na seguinte função de probabilidade:

$$P(X = x) = \binom{r + x - 1}{x} \left(\frac{\beta}{\beta + 1}\right)^r \left(\frac{1}{\beta + 1}\right)^x \qquad x = 0,1,\cdots$$

A distribuição resultante é, então, uma Binomial Negativa com parâmetros r e $p = \left(\frac{\beta}{\beta + 1}\right)$.

Nas metodologias utilizadas para avaliar a existência de expressão diferencial entre as condições, como por exemplo, no edgeR (ROBINSON et al., 2010) e DESeq (ANDERS e HUBER, 2010), é realizada uma reparametrização no modelo, igualando-se o valor esperado a μ, como na expressão:

$$E(X) = \frac{r(1-p)}{p} = \mu$$

E após seu desenvolvimento, pode-se mostrar que:

$$p = \frac{1}{1+\phi\mu}$$

Em que $\phi = \frac{1}{r}$. Substituindo esse resultado em $V(X) = \frac{r(1-p)}{p^2}$, temos que:

$$V(X) = \frac{r(1-p)}{p^2} = \mu + \phi\mu^2$$

Com essa reparametrização, nota-se que o modelo Binomial Negativo é mais flexível que o Poisson, pois possibilita a modelagem da variabilidade por meio da média. A função de probabilidade do modelo reparametrizado é dada por:

$$P(X = x) = \binom{r+x-1}{x}\left(\frac{1}{1+\phi\mu}\right)^{\phi^{-1}}\left(1-\frac{1}{1+\phi\mu}\right)^{x} \qquad x = 0,1,\cdots$$

2.3.4 Correção para testes múltiplos

Um dos principais objetivos em análises de RNA-Seq é a identificação de genes que se expressam diferencialmente e cujo nível de expressão está associado a uma resposta ou variável de interesse. Como a detecção de genes diferencialmente expressos envolve a realização de um grande número de testes estatísticos, tendo várias hipóteses testadas, a probabilidade conjunta de que o erro tipo I seja cometido aumenta expressivamente. O erro tipo I, também denominado de falso positivo (CASELLA e BERGER, 2010), é o erro que se comete ao rejeitar a hipótese nula quando a mesma é verdadeira (BENJAMINI & HOCHBERG, 1995). No caso de análises de RNA-Seq, ele ocorre ao afirmar que um gene se expressa diferencialmente, quando na realidade isso não acontece.

A correção convencional do p-valor para a multiplicidade, visa controlar a taxa de erro do tipo I, mas muitas vezes é conservador devido a grande quantidade de testes que é realizada nesse tipo de experimento. Assim, ao avaliar a significância estatística das detecções, tornou-se uma prática comum utilizar a FDR (False Discovery Rate) para controlar a taxa de falsas descobertas das detecções, ou seja, a proporção esperada de falsos positivos entre todas as detecções, e corrigir os p-valaores utilizando, por exemplo, o método de Hochberg-Benjamini (BENJAMINI e HOCHBERG, 1995).

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 de significância conjunta aumenta consideravelmente, justificando a necessidade da correção dos testes múltiplos.

2.3.4.1 Proteção de Bonferroni

O objetivo da utilização da proteção de Bonferroni é prover 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 e CRUZ, 2008):

$$\alpha = \frac{\alpha^T}{m}$$

em que: α é o nível de significância de cada teste;

α^T é o nível de significância conjunta do experimento e

m é a quantidade de testes realizados

Devido à grande quantidade de genes estudados, ao ser utilizada em experimentos de RNA-Seq, a correção de Bonferroni torna-se muito conservadora, diminuindo sobremaneira 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, tendo em vista 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 expressões 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.

2.3.4.2 FDR - False Discovery Rate

Benjamini e 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, a FDR é calculada após a realização dos múltiplos testes de hipóteses, a partir das informações presentes nos dados.

Um procedimento bastante utilizado para calcular a FDR é o Linear Step-Up proposto por Benjamim & Hochberg (1995). Tal procedimento ordena os p-valores $p_{(1)} \leq \cdots \leq p_{(m)}$ resultantes das m hipóteses $H_1, \cdots, H_m$, testadas de forma simultânea. Sejam $p_{(1)}, p_{(2)}, \cdots, p_{(m)}$ os p-valores ordenados, define-se:

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

Assim, 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 $P_{(i)}$.

2.4 Avaliação dos genes diferencialmente expressos

A fim de identificar genes diferencialmente expressos existem várias metodologias implementadas em pacotes do software R (R Development Core Team, 2012). Como exemplo, apresentam destaque na literatura as metodologias EdgeR (ROBINSON et al., 2010), DESeq (ANDERS & HUBER, 2010), DEGSeq (WANG et al., 2010) e baySeq (HARDCASTLE e KELLY, 2010).

2.4.1 DESeq

O DESeq é um pacote desenvolvido para analisar dados de contagem a partir de ensaios de sequenciamento de alto rendimento, tais como o RNA-Seq e proporciona formas para testar a expressão diferencial por meio da distribuição binomial negativa.

Os dados de contagem apresentam uma distribuição de Poisson. Porém, 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 e os testes baseados na distribuição de Poisson ignoram a variação de amostragem biológica, ou seja, para dados que apresentam

superdispersão, análises baseadas em Poisson tenderão a uma alta taxa de falsos positivos resultantes de uma subestimação do erro de amostragem (KVAM e LIU, 2012).

Quando existem repetições biológicas (amostragem de mais de um indivíduo em cada tratamento), os dados de RNA-Seq podem apresentar maior variabilidade, ou seja, a variância é capaz de ultrapassar o valor da média consideravelmente para muitos genes, que é definido em literatura como problema de superdispersão (KVAM e LIU, 2012).

Uma solução para o problema mencionado acima, é assumir que os dados seguem uma distribuição binomial negativa, que é um prolongamento natural da distribuição de Poisson, já que é geralmente utilizada para lidar com o excesso de dispersão. Assim, o DESeq aplica um modelo baseado na distribuição binomial negativa com o intuito de controlar esse excesso de variabilidade na amostra, permitindo a detecção dos genes diferencialmente expressos e modelando a variação biológica corretamente.

2.4.1.1 Descrição do modelo utilizado

Considere que o número de reads do gene i na j-ésima amostra (biblioteca), segue uma distribuição binomial negativa (ANDERS & HUBER, 2010).

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

em que K_{ij} é o número de reads i, na biblioteca j, μ_{ij} é a média e σ_{ij}^2 é a variância.

Na prática, tais parâmetros são desconhecidos e devem ser estimados por meio do conjunto de observações. Geralmente, estudos de expressão diferencial apresentam um pequeno número de repetições e desse modo, visando a obtenção de estimativas razoáveis, a metodologia implementada no pacote DESeq (ANDERS & HUBER, 2010) baseia-se em três suposições:

1. O parâmetro média, μ_{ij}, ou seja, o valor esperado do número de reads para o gene i na j-ésima amostra, é o produto de um valor dependente da condição do gene $q_{i,\rho(j)}$, e do fator do tamanho da biblioteca s_j, isto é,

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

em que ρ_j denota a condição experimental da j-ésima amostra, $q_{i,\rho(j)}$ são valores proporcionais aos esperados da verdadeira concentração de fragmentos do gene i, sob a

condição experimental j. O fator de tamanho, s_j, representa a cobertura da biblioteca (amostra) j.

2. A variância σ_{ij}^2 é dada pela soma da média e da variância bruta (VB) dos dados,

$$\sigma_{ij}^2 = \mu_{ij} + \underbrace{s_j^2 \nu_{i,\rho(j)}}_{VB}$$

3. Assume-se que o parâmetro $\nu_{i,\rho(j)}$ da variância bruta é uma função suave de $q_{i,\rho(j)}$, ou seja,

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

A última suposição nos permite utilizar todos os dados de genes com expressão similar para estimar a variância.

2.4.1.2 Estimação dos parâmetros do modelo

Para o ajuste do modelo aos dados, primeiramente esses devem ser organizados em uma tabela de contagem de ordem g x m, em que g representa o número de genes avaliados e m o número de bibliotecas (amostras). Como exemplo, considera-se um conjunto de dados com g genes e 2 bibliotecas:

Gene	A_1	A_2
1	K_{11}	K_{12}
2	K_{21}	K_{22}
3	K_{31}	K_{32}
…	…	…
g	K_{g1}	K_{g2}

O modelo adotado possui três conjuntos de parâmetros:

- m fatores de tamanho s_j (soma dos reads da amostra j);
- para cada condição experimental ρ, g parâmetros $q_{i\rho}$;
- Funções suaves ν_ρ; para cada condição ρ, ν_ρ modelos de dependência da variância em relação à média esperada $q_{i\rho}$.

A fim de processar os dados, primeiramente deve ser realizada a normalização, para tornar os valores de contagens de diferentes amostras (bibliotecas), comparáveis, tendo em vista que uma quantidade maior de material genético está associada a uma maior quantidade de reads. Para isso, utiliza-se o fator de tamanho s_j. O estimador do fator de tamanho é a mediana de razão das contagens observadas.

$$\hat{s}_j = \text{median}\left[\frac{k_{ij}}{\left(\prod_{\nu=1}^{m} k_{i\nu}\right)^{\frac{1}{m}}}\right],$$

em que k = contagens observadas e m = número de repetições

Para estimar $q_{i\rho}$, utiliza-se a média do número de contagens das amostras j correspondendo à condição experimental ρ, transformada para a escala comum por meio do fator de tamanho:

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

em que m_ρ é o número de repetições da condição ρ.

Para estimar a função suave ν_ρ, primeiramente devem ser calculadas as variâncias amostrais em escala comum. Para obter essas quantidades, são necessários os seguintes elementos:

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

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

A quantidade $z_{i\rho}$ é a correção do viés do estimador da variância bruta. De acordo com Anders & Huber (2010), $\omega_{i\rho} - z_{i\rho}$ é um estimador não viesado da variância bruta $s_j^2 \nu_{i,\rho(j)}$. Apesar da utilidade, para os casos em que o número de repetições é pequeno, como nos experimentos de RNA-Seq, a quantidade $\omega_{i\rho}$ é altamente variável, portanto $\omega_{i\rho} - z_{i\rho}$ não é um estimador interessante para realizar inferências. Uma alternativa proposta por Anders & Huber (2010) é a utilização de um modelo de regressão não paramétrico considerando como

variável dependente $\omega_{i\rho}$ e independente $\hat{q}_{i\rho}$ para obter a função suave $\omega_\rho(q)$ e estimar a variância bruta por meio de:

$$\hat{\nu}_\rho\left(\hat{q}_{i\rho}\right) = \omega_\rho\left(\hat{q}_{i\rho}\right) - z_{i\rho}$$

2.4.1.3 Avaliação da expressão diferencial por meio do DEseq para duas condições

O teste exato de Fisher é utilizado em situações onde não se observam repetições biológicas, ou seja, situações com um único indivíduo por grupo de tratamento. Nesses casos não é possível estimar a variabilidade dentro do grupo de tratamento, então a análise deve prosseguir sem qualquer informação sobre a variação biológica dentro do grupo.

Para avaliar a expressão diferencial, a literatura apresenta como solução o teste "exato de Fisher" (Fisher, 1934). Tal análise é geralmente realizada gene a gene, organizando os dados em uma tabela 2x2.

Para ilustrar o procedimento, considere o conjunto de observações:

Gene	G_1	G_2
1	K_{11}	K_{12}
2	K_{21}	K_{22}
3	K_{31}	K_{32}
…	…	…
g	K_{g1}	K_{g2}

Considerando-se a avaliação do gene 1, constrói-se a tabela de dupla entrada:

	C_1	C_2	Total
Gene 1	n_{11}	n_{12}	$N_{1.}$
Genes restantes	n_{21}	n_{22}	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	$N_{..}$

Nessa tabela, $n_{11}=K_{11}$, $n_{12}=K_{12}$, $n_{21} = \sum_{i=2}^{g} K_{i1}$ e $n_{21} = \sum_{i=2}^{g} K_{i2}$

O teste verifica se a proporção de contagens para um dado gene nas duas condições é a mesma que para os demais, ou seja:

$$\frac{\pi_{11}}{\pi_{12}} = \frac{\pi_{21}}{\pi_{22}}$$

Essa hipótese em geral é escrita em termos da razão de chances, como se segue:

$$\begin{cases} H_0 : \theta = \dfrac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1; \\ H_1 : \theta \neq 1. \end{cases}$$

Assim, no caso de $N_{1.}$ reads mapeadas para o gene 1 e $N_{2.}$ reads para o restante dos genes, se forem extraídos aleatoriamente $N_{1.}$ reads do conjunto total de reads avaliados, questiona-se qual a probabilidade de se observar um resultado ao menos igual a n_{11} reads para o gene 1. Se essa probabilidade é pequena, então a classificação das colunas da tabela, ou seja, da condição experimental afetou a amostragem. Especificamente, o gene 1 é diferencialmente expresso entre as condições 1 e 2. O cálculo desta probabilidade é realizado por meio da distribuição hipergeométrica (AGRESTI, 1990), ou seja,

$$p = \frac{\binom{N_{1.}}{n_{11}}\binom{N_{2.}}{n_{21}}}{\binom{N_{..}}{n_{11}+n_{21}}}$$

O p-valor é obtido somando-se todas as probabilidades de se observar um número k maior ou igual a n_{11}, ou seja:

$$p(\text{reads} \geq n_{11}) = \sum_{k=n_{11}}^{N_{1.}} \frac{\binom{k+n_{12}}{k}\binom{N_{2.}}{n_{21}}}{\binom{N_{..}}{k+n_{21}}}$$

Para avaliar a existência de expressão diferencial por meio do teste implementado no DESeq, considere m_A amostras repetidas para a condição A e m_B amostras repetidas para a condição B. Para cada gene, a hipótese a ser avaliada é dada por:

$$H_0 : q_{iA} = q_{iB}$$

em que q_{iA} é a média da contagem do gene i na condição A e q_{iB} é a média da contagem do gene i na condição B

O teste utilizado para avaliar tal hipótese é um teste condicional bem semelhante ao teste exato de Fisher apresentado anteriormente. A principal diferença se deve ao uso da

distribuição de probabilidade adotada para o cálculo do p-valor. Diferentemente do teste exato de Fisher, o qual se baseia na distribuição hipergeométrica, o teste aqui apresentado baseia-se na distribuição binomial negativa possibilitando inserir no processo de decisão informações a respeito da variação biológica dentro do grupo.

Sob a hipótese nula pode-se calcular as probabilidades dos eventos $K_{iA} = a$ e $K_{iB} = b$ para quaisquer pares de números (a,b). O valor p de um par observado é dado pela soma das probabilidades menores ou iguais a $p(k_{iA}, k_{iB})$ dada a soma total k_{is},

$$p_i = \frac{\sum_{p(a,b) \leq p(k_{iA}, k_{iB})} p(a,b)}{\sum_{a+b=k_{is}} p(a,b)}$$

em que as variáveis a e b assumem valores entre $0, \cdots, k_{is}$.

Para o cálculo de p(a,b), deve-se assumir que, sob a hipótese nula, as amostras são independentes, ou seja, $p(a,b) = P(K_{iA} = a)P(K_{iB} = b)$. A parametrização da distribuição binomial negativa apresentada em Anders & Huber (2010) é dada por:

$$P(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k,$$

em que p e r podem ser parametrizados em termos da média μ e da variância σ^2 por meio de :

$$p = \frac{\mu}{\sigma^2} \quad ; \quad r = \frac{\mu^2}{\sigma^2 - \mu}.$$

2.4.2 baySeq

O baySeq utiliza uma abordagem inferencial bayesiana empírica para estimar a probabilidade a posteriori de cada gene se expressar diferencialmente sob determinadas condições, as quais são representadas por meio de modelos (SONESON E DELORENZI, 2013). Nesta abordagem, como em diversas outras, é assumido, pelas razões discutidas anteriormente, que o número de reads segue uma distribuição binomial negativa.

2.4.2.1. Definição dos modelos

Considere o caso de análise mais simples, no qual o interesse seja comparar duas condições experimentais, por exemplo, as condições A e B. Considere também que para cada

condição temos duas repetições biológicas, ou seja, tem-se no total quatro bibliotecas denotadas por A_1, A_2, B_1, B_2. De acordo com Hardcastle e Kelly (2010) é razoável supor que a maioria dos genes não é afetada pelas condições experimentais e, desta forma, os dados (reads) possuem os mesmos parâmetros nestas duas condições. No entanto, alguns genes podem apresentar expressão diferencial entre as diferentes condições experimentais e assim, os valores dos parâmetros referentes as amostras da condição A (A_1 e A2) serão diferentes daqueles provenientes das amostras referentes a condição B (B_1 e B_2). Assim, como apresentado em Hardcastle & Kelly (2010) podemos definir dois modelos. O primeiro, no qual é assumido que não existe expressão diferencial entre as condições (amostras) e, consequentemente os parâmetros são os mesmo nas amostras, o conjunto de dados é dado por todas as amostras conjuntamente, ou seja, $\{A_1, A_2, B_1, B_2\}$. No segundo modelo, em que pressupõe-se expressão diferencial entre as condições A e B (as amostras são definidas por $\{A_1, A_2\}$ e $\{B_1, B_2\}$, ou seja, os parâmetros diferem nas duas condições experimentais. Modelos de análise mais complexos, como por exemplo, com três condições experimentais, podem ser vistos em Hardcastle e Kelly (2010).

2.4.2.2 Abordagem bayesiana para inferência sobre expressão diferencial

Suponha um conjunto de dados de contagens com n amostras, $A = \{A_1, \ldots, A_n\}$, tais que os dados observados para um determinado gene é dado pelo seguinte conjunto:

$$D_c = \{(u_{1c}, \cdots, u_{nc}), (l_1, \cdots, l_n)\},$$

em que u_{ic} é a contagem do c-ésimo "gene" avaliado na i-ésima amostra (biblioteca). Para cada amostra A_i, define-se também o tamanho da biblioteca como um fator de escala l_i.

Considere então, um modelo M que pode ser definido pelos seguintes conjuntos $\{E_1, \cdots, E_m\}$. Assim, como apresentado na seção anterior, se as amostras A_i e A_j pertencerem a um mesmo conjunto, digamos E_q, estas possuem os mesmos parâmetros, ou seja, $\theta_{A_1} = \theta_{A_2} = \theta_q$. Assim, os dados associados ao conjunto E_q para um determinado gene podem ser representados por:

$$D_{qc} = \{(u_{ic} : A_i \in E_q), (l_i : A_i \in E_q)\}.$$

Sob o enfoque bayesiano, visando avaliar expressão diferencial entre as diferentes condições experimentais por meio de um modelo M para os dados, a quantidade de interesse para cada gene c, é a probabilidade à posteriori do modelo M dado os dados D_c, ou seja,

$$P(M \mid D_c) = \frac{P(D_c \mid M)P(M)}{P(D_c)}.$$

Para a obtenção dessa quantidade será necessário obter a probabilidade marginal das observações dado o modelo, ou seja, $P(D_c \mid M)$, estimar a probabilidade a priori de cada modelo e obter o fator de escala $P(D_c)$. A seguir, será apresentada de maneira sucinta a obtenção de cada uma destas quantidades.

2.4.2.2.1 Obtenção de $P(D_c \mid M)$

Segundo Hardcastle & Kelly (2010), $P(D_c \mid M)$ pode ser calculada por meio da verossimilhança marginal dada por:

$$P(D_c \mid M) = \int P(D_c \mid K, M)P(K \mid M)dK,$$

em que $P(D_c \mid K, M)$ é a distribuição dos dados condicionada ao modelo e ao conjunto de parâmetros $(K = \{\theta_1, \ldots, \theta_m\})$ e $P(K \mid M)$ é a distribuição dos parâmetros condicionada ao modelo avaliado.

Existem inúmeras distribuições que podem ser utilizadas para $D_c \mid K, M$ e K|M. Como discutido em Hardcastle & Kelly (2010), uma abordagem natural para obtenção de $P(D_c \mid K, M)$ seria assumir que os dados possuem distribuição Poisson e que $P(K \mid M)$, tem distribuição Gama. Porém, Robinson e Smyth (2008) afirmam que essa modelagem não é adequada quando se leva em consideração a variabilidade extra, introduzida pelas repetições biológicas. Assim, visando explicar a tal variabilidade, pode-se supor que os dados possuem distribuição binomial negativa a qual é indicada a fenômenos que apresentam super dispersão. Além disso, Lu et al (2005) mostraram em dados simulados que a suposição de uma distribuição binomial negativa pode ser robusta, mesmo que os dados não possuam verdadeiramente distribuição binomial negativa.

Visando trabalhar com os dados originais, ou seja, levando em consideração os tamanhos da biblioteca, Hardcastle & Kelly (2010) fizeram uso de métodos numéricos para obtenção destas quantidades.

Considere que a amostra A_i pertencente ao conjunto E_q com o tamanho da biblioteca l_i. Assim, considerando que u_{ic} (contagem do c-ésimo gene - D_{qc}) segue uma distribuição binomial negativa com média $\mu_q l_i$ e variância φ_q, em que $\theta_q = (\mu_q, \varphi_q)$ temos,

considerando a parametrização apresentada em 2.3.3, a seguinte distribuição de probabilidades:

$$P\left(u_{ic};l_i,\phi_q,\mu_q\right)=\frac{\Gamma\left(u_{ic}+\phi_q^{-1}\right)}{\Gamma\left(\phi_q^{-1}\right)u_{ic}!}\left(\frac{1}{1+l_i\mu_q\phi_q}\right)^{\phi_q^{-1}}\left(\frac{l_i\mu_q}{\phi_q^{-1}+l_i\mu_q}\right)^{u_{ic}}.$$

Infelizmente, nesse caso não é possível encontrar uma conjugação óbvia como no modelo Poisson Gama. Assim, visando estimar $P(D_c \mid M)$, numericamente, é necessário definir uma distribuição empírica para K. Para tanto, Hardcastle & Kelly (2010) assumiram que $\theta_q \in K$ e são independentes em relação à q, ou seja,

$$\begin{aligned}P(D_c \mid M)&=\int P(D_c \mid K,M)P(K \mid M)dK\\&=\prod_q\int P(D_{qc} \mid \theta_q)P(\theta_q)d\theta_q.\end{aligned}$$

Essa suposição reduz a dimensão da integral, melhorando a precisão da de sua aproximação. Em seguida, supõe-se que para cada $\theta_q \in K$, há um conjunto de valores Θ_q (espaço paramétrico) que são amostrados a partir da distribuição de θ_q. Com tais suposições, conforme apresentado em Evans e Swartz (1995) podemos derivar a aproximação:

$$P(D_c \mid M)=\prod_q\frac{1}{|\Theta_q|}\sum_{\Theta_q}\left[\prod_{\{i:A_i\in E_q\}}\frac{\Gamma\left(u_{ic}+\phi_q^{-1}\right)}{\Gamma\left(\phi_q^{-1}\right)u_{ic}!}\left(\frac{1}{1+l_i\mu_q\phi_q}\right)^{\phi_q^{-1}}\left(\frac{l_i\mu_q}{\phi_q^{-1}+l_i\mu_q}\right)^{u_i}\right].$$

A partir de então, para obter tal aproximação é necessário obter o conjunto Θ_q com base nos dados.

2.4.2.2.2 Obtenção da distribuição empírica sobre K

A obtenção da distribuição empírica de K pode ser realizada por meio da análise de todo o conjunto de dados. Assim, para cada conjunto de amostras, E_q, devemos encontrar uma estimativa da média e da dispersão da distribuição de probabilidades de D_{qc}. Após a obtenção de tais estimativas para um grande número de genes teremos nossa amostra Θ_q.

Segundo Hardcastle & Kelly (2010) o principal problema nessa abordagem se refere à obtenção das estimativas de dispersão, visto que, se os dados de um determinado gene apresentam expressão diferencial e o modelo que sob teste presume que não existe expressão

diferencial, a dispersão irá ser superestimada para esse gene. Para contornar tal problema os autores sugerem considerar a estrutura de repetição dos dados para estimar corretamente as dispersões. A estrutura de repetição é definida considerando-se os conjuntos $\{F_1, \cdots, F_s\}$, em que $i, j \in F_r$ se e somente se a amostra e A_j é repetição de A_i.

Dada essa estrutura para os dados pode-se estimar a dispersão para um determinado gene, D_c, por métodos de quasi-verossimilhança apresentados em Nelder (2000) (HARDCASTLE & KELLY, 2010). Robinson & Smyth (2008) mostraram que os métodos de quasi-verossimilhança fornecem bons resultados quando utilizados para a estimação de um único gene. A estimação da dispersão ocorre por meio de um processo iterativo em que primeiramente define-se:

$$\hat{\mu}_{ic} = \left\langle \left\{ \frac{u_{ic}}{l_i} : i \in F_r \right\} \right\rangle,$$

e, em seguida escolhe-se ϕ_c, tal que

$$2\sum_r \sum_{i \in F_r} \left\{ u_{ic} \log\left[\frac{u_{ic}}{l_i \hat{\mu}_{ic}} \right] - \left(u_{ic} + \phi_c^{-1}\right) \log\left[\frac{u_{ic} + \phi_c^{-1}}{l_i \hat{\mu}_{ic} + \phi_c^{-1}} \right] \right\} = n - 1.$$

Tomando esse valor para ϕ_c pode-se então reestimar os valores $\hat{\mu}_{ic}$ pelo método da máxima verossimilhança, escolhendo-se os valores para $\hat{\mu}_{ic}$ que, para cada r, maximizam a seguinte verossimilhança:

$$P\left(\{u_{ic} : i \in F_r\}; l_i : i \in F_r, \phi_c, \hat{\mu}_{ic}\right) = \prod_{i \in F_r} \frac{\Gamma\left(u_{ic} + \phi_c^{-1}\right)}{\Gamma\left(\phi_c^{-1}\right) u_{ic}!} \left(\frac{1}{1 + l_i \hat{\mu}_{ic} \phi_c} \right)^{\phi_c^{-1}} \left(\frac{l_i \hat{\mu}_{ic}}{\phi_c^{-1} + l_i \hat{\mu}_{ic}} \right)^{u_{ic}}.$$

Esse processo é repetido até que as estimativas ϕ_c e $\hat{\mu}_{ic}$ atinjam a convergência. Assim, após a obtenção de um valor para ϕ_c pelo método descrito anteriormente, precisamos estimar a média da distribuição dos dados D_{qc} para todos os conjuntos de amostras E_q. Para tanto, fixamos o valor obtido de ϕ_c e estimamos, para cada q, a média μ_{qc} maximizando a seguinte verossimilhança:

$$P\left(D_{qc}, \phi_c, \mu_{qc}\right) = \prod_{i : A_i \in E_q} \frac{\Gamma\left(u_{ic} + \phi_c^{-1}\right)}{\Gamma\left(\phi_c^{-1}\right) u_{ic}!} \left(\frac{1}{1 + l_i \mu_{qc} \phi_c} \right)^{\phi_c^{-1}} \left(\frac{l_i \mu_{qc}}{\phi_c^{-1} + l_i \mu_{qc}} \right)^{u_{ic}}.$$

O conjunto $\Theta_q = \{(\mu_{qc}, \varphi_c)\}$ é obtido repetindo-se esse processo para vários h, possibilitando assim o cálculo de

$$P(D_c \mid M) = \prod_q \frac{1}{|\Theta_q|} \sum_{\Theta_q} \left[\prod_{\{i:A_i \in E_q\}} \frac{\Gamma\left(u_{ic} + \phi_q^{-1}\right)}{\Gamma\left(\phi_q^{-1}\right) u_{ic}!} \left(\frac{1}{1 + l_i \mu_q \phi_q} \right)^{\phi_q^{-1}} \left(\frac{l_i \mu_q}{\phi_q^{-1} + l_i \mu_q} \right)^{u_i} \right].$$

2.4.2.2.3 Estimação da probabilidade a priori de cada modelo

Uma série de opções está disponível quando se consideram as probabilidades a priori de cada modelo P(M) exigido para o cálculo de $P(M \mid D_c)$. Na proposta de Hardcastle e Kelly (2010), inicialmente deve-se escolher um valor p, para que o mesmo seja utilizado como probabilidade a priori do modelo M para o cálculo da probabilidade a posteriori, $P(M \mid D_c)$, do c-ésimo "gene". A partir desse valor, pode-se encontrar uma nova estimativa para a probabilidade a priori do modelo M dada por $p' = \langle P(M \mid D_c) \rangle_c$. Esse processo deve ser repetido até que o valor de p apresente convergência.

2.4.2.3 O fator de escala $P(D_c)$

Finalmente, é necessário obter o fator de escala, $P(D_c)$, da equação $P(M \mid D_c)$. Uma vez que o número de possíveis modelos de M em A é finito, o fator de escala $P(D_c)$ pode ser determinado pela soma de todos os possíveis M, dada as prioris apropriadas P(M).

2.4.2.4 Avaliação das hipóteses

A avaliação das hipóteses é realizada por meio das probabilidades a posteriori, ou seja, se o modelo que contempla expressão diferencial apresentar maior valor de probabilidade quando comparado ao modelo que considera que as amostras tenham os mesmos valores paramétricos, deve-se rejeitar a hipótese da não existência da expressão diferencial. Caso contrário, não se deve rejeitar a hipótese de igualdade dos parâmetros.

2.5 Curva ROC (Receiver Operating Characteristic)

A curva ROC (FAWCETT, 2006) é utilizada para avaliar o desempenho de classificadores binários. Dados uma instância e um classificador que rotula em positivo ou negativo, existem quatro configurações possíveis. Se a instância for corretamente classificada como positivo, ela é um verdadeiro positivo. Se for incorretamente classificada como

positivo, ela é um falso positivo. Caso a instância seja classificada como negativo e de fato o seja, ela é um verdadeiro negativo. Por fim, se a instância for incorretamente classificada como negativo, ela é um falso negativo. As quatro configurações aqui expostas podem ser resumidas na tabela de contingência 2 x 2 exibida a seguir:

Resultado do teste	**Valor Verdadeiro**	
	positivo	negativo
positivo	VP Verdadeiro Positivo	FP Falso Positivo
negativo	FN Falso Negativo	VN Verdadeiro Negativo

Define-se:

- Taxa de verdadeiro positivo: $TVP = \frac{VP}{\text{Total de positivos}} = \frac{VP}{VP + FN}$
- Taxa de falso positivo: $TFP = \frac{FP}{\text{Total de negativos}} = \frac{FP}{FP + VN}$

A curva ROC é desenhada na grade da taxa de falso positivo versus a taxa de verdadeiro positivo. Cada ponto da curva está associado a um limiar para a classificação em positivo e negativo (supondo que o resultado devolvido pelo classificador é um valor utilizado para discriminar as duas classes) (ZWEIG e CAMPBELL, 1993).

Uma curva ROC que passa pela coordenada (0, 1) representa o melhor desempenho possível de um classificador, pois indica uma taxa de verdadeiro positivo de 100% (isto é, ausência de falsos negativos), e uma taxa de falso positivo de 0% (VAN DER SCHOUW et al., 1992).

Um classificador que rotula a instância em positivo ou negativo aleatoriamente, teria uma curva ROC na linha diagonal, com os extremos nas coordenadas (0, 0) e (1, 0).

Assim, um classificador com desempenho melhor do que uma rotulação aleatória deve ter uma curva ROC correspondente acima da linha diagonal. O desenho da curva nos permite, então, comparar diferentes classificadores.

A área abaixo de uma curva ROC, AUC (Area Under Curve), é a sensibilidade média dentre todos os valores de especificidade possíveis, e pode ser usada como medida de para estimar o desempenho da predição para diferentes limites (LIU et al, 2006). Sendo a AUC uma porção de um gráfico quadrado, seu valor fica entre 0 e 1, de modo que se um classificador possui algum poder discriminativo deve ter sua AUC maior que 0,5, que é o valor encontrado para um classificador que realiza predições aleatórias (FAWCETT, 2006). A área sob a curva pode também ser interpretada como a probabilidade de que um classificador dê uma pontuação maior para instâncias positivas do que para instâncias negativas (FAWCETT, 2006).

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. **Categorical data analysis**. New York: Wiley, 1990

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

AUER, P. L.; DOERGE, R. W. A two-stage poisson model for testing RNA-Seq data**. Stat Appl Gen Mol Biol**, 10:Article 26. 2011.

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

BENJAMINI, Y.; HOCHBERGH,Y. **Controlling the false discovery rate - a practical and powerful approach to multiple testing**. JRStat Soc BMethodol 1995;57:289–300.

BULLARD, J., PURDOM, E., HANSEN, K., DURINCK, S. & DUDOIT, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. **BMC Bioinformatics** 11, 94. 2010.

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

DANTAS, D. O. **Uma técnica automática baseada em morfologia matemática para medida de sinal em imagens de cDNA.** Dissertação de mestrado (ciência da computação). São Paulo: Universidade de São Paulo, 2004.

DILLIES M.A., RAU A., AUBERT J. **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. Brief Bioinform 2013;14:671–683.

EVANS, M.; SWARTZ, T. Métodos de aproximação integrais em estatística, com ênfase especial em bayesian integração problemas. **Statistical Science**, 10 (3): 254-272, 1995.

FAWCETT, T. **An introduction to ROC analysis**. Pattern Recognition Letters, 27:861–874, 2006.

FISHER, R.A. **Statistical Methods for Research Workers**. 5th Edition, Edinburgh: Oliver and Boyd. 1934.

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**, 5:R80. 2004.

GONÇALVES, J. C. **Influência do número de repetições na identificação de genes diferencialmente expressos em experimentos de RNA-Seq.** Dissertação (Mestrado em Estatística Aplicada e Biometria) Universidade Federal de Viçosa, 2013.

HARDCASTLE, T. J.; KELLY, K. A. Bayseq: Empirical bayseian methods for identifying differential expression in sequence count data. **BMC Bioinformatics**, 11 R: 106, 2010.

HINDE, J.;DEMÉTRIO, C. **Overdispersion: models and estimation.Computational** Statistics and Data Analysis, 1998, 27, 151–170

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.

LANGMEAD B, HANSEN KD, LEEK JT. Cloud-scale RNA-Sequencing differential expression analysis with Myrna. **Genome Biology**. 2010;11(8):R83

LI, J.; TIBSHIRANI, R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. **Statistical Methods in Medical Research**, v. 22, p. 519-536, 2013.

LIU J., GOUGH, J., ROST, B. (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines. **PLoS Genet** 2: e29. doi:10.1371/journal.pgen.0020029.

LOVE, M. I., W. HUBER, S. ANDERS: Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. **bioRxiv** (2014). doi:10.1101/002832.

LU, C., TEJ, S.S., LUO, S., HAUDENSCHILD, C.D., MEYERS, B.C., GREEN, P.J.(2005) Elucidation of the small RNA component of the transcriptome. **Science** 309:1567–1569.

MADEIRA, A. P. C.; CHAVES, L. M.; SOUZA, D. J. A mathematical validation for an algorithm that simulates mixture of distributions. Rev. Bras. Biom., São Paulo, v.27, n.4, p.603-612, 2009.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. (1974). **Introduction to the theory of statistics** (3rd ed.). Tokyo: McGraw-Hill.

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.

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

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.

NAGALAKSHMI U., et al. 2008. The Transcriptional landscape of the yeast genome defined by RNA sequencing. **Science** 320 (5881):1344-9

NELDER, J. **Quasi-likelihood and psuedo-likelihood are not the same thing.** Journal of Applied Statistics, 27(8):1007-1011. 2000.

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

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. **EdgeR: a bioconductor package for differential expression analysis of digital gene expression data.** Bioinformatics, 26:139–140. 2010.

ROBINSON, M. D.; SMYTH, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. **Biostatistics**.9, 2, pp. 321–332, 2008.

ROBLES, J. A.; QURESHI, S. E.; STEPHEN, S. J.; WILSON, S. R.; BURDEN, C. J.;

TAYLOR, J. M. - Efficient experimental design and analysis strategies for the detection of differential expression using RNA Sequencing. **BMC Genomics**, v. 13, p. 484, 2012.

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.

SMYTH, G. K., THORNE, N. P., AND WETTENHALL, J. (2003). **Limma: Linear Models for Microarray Data User's Guide.** Software manual available from http://www.bioconductor.org.

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

SONESON, C.; DELORENZI, M. A comparison of statistical methods for detecting diferentially expressed genes from RNA-Seq data**. BMC Bioinformatics**, 14:91, 2013.

TARAZONA, S.; GARCÍA, A. F.; DOPAZO, J.; FERRER, A.; CONESA, A. **Differential expression in RNA-Seq: a matter of depth**. Genome Res, 21:2213–2223. 2011.

VAN DER SCHOUW, Y.; VERBEEK A.; RUIJS, J (1992) **ROC curves for the initial assessment of new diagnostic tests**. Family Practice 9: 506–511 [PubMed].

WANG, L.; FENG, Z.; WANG, X.; ZHAN G. **Degseq: an R package for identifying differentially expressed genes from RNA-Seq data.** Bioinformatics, 26, 136-138, 2010.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for Transcriptomics. **Nat. Reviews Genetics** (10) 57-63, 2009.

ZWEIG, M.H.; CAMPBELL, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. **Clin Chem** 1993;39:561–77.

CAPÍTULO 1

Número de repetições na identificação de genes diferencialmente expressos em experimentos de RNA-Seq

Resumo

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 biológicos responsáveis pelas transformações moleculares. Esta análise tem sido realizada pela obtenção do RNA-Seq (transcriptoma) produzido pelas diferentes tecnologias de sequenciamentos de nova geração (NGS), a partir do cDNA derivado de RNA, que tem produzido milhões de leituras representativas dos genes expressos. Porém, como o custo do sequenciamento por essas tecnologias era alto, a realização de experimentos com poucas ou nenhuma repetição tornou-se comum. Assim, tornou-se necessária a descoberta e o aprimoramento de metodologias estatísticas eficientes para a otimização das análises dos dados gerados nestas plataformas de sequenciamento. Esse trabalho avaliou as metodologias estatísticas DESeq e Bayseq com a finalidade de verificar a influência do número de repetições na identificação de genes diferencialmente expressos em experimentos de RNA-Seq. Para tanto, utilizou-se um conjunto de dados simulados em diferentes cenários. Tais cenários foram constituídos por duas condições, compostas por 1000 genes obtidos por RNA-Seq com diferentes perfis de expressão gênica. O número de repetições variou entre os diferentes cenários, sendo estes com cinco repetições, três repetições, duas repetições e sem repetição. Em um primeiro momento, os dados foram analisados pelos dois métodos separadamente, comparando-se o efeito do número de repetições dentro de cada um. Em seguida, foi realizada a comparação entre os métodos, levando em conta também o número de repetições em cada cenário. De acordo com os resultados gerados não podemos afirmar que um método, entre os avaliados, é ótimo em todas as circunstâncias, pois depende das condições experimentais. No entanto, sob as condições utilizadas no desenvolver do experimento, o método abordado pelo baySeq foi o que apresentou um melhor desempenho, nas combinações ocorridas entre os métodos e os tipos de genes analisados, uma vez que obteve maior capacidade de identificar genes diferencialmente expressos.

Palavras-chave: RNA-Seq, transcriptoma, repetições, simulação.

1. Introdução

O método de sequenciamento de cDNA, RNA-Seq, tornou-se em pouco tempo a ferramenta mais indicada para estudos de transcriptoma, visto que possui vantagens em relação as técnicas anteriores (microarrays, SAGE), como por exemplo, possui uma maior sensibilidade quando comparada à técnica de microarrays (SEYEDNASROLLAH, 2013).

Em muitos estudos de RNA-Seq o problema central de pesquisa é a identificação de genes diferencialmente expressos entre dois ou mais grupos de amostras distintas, as quais se caracterizam por diferentes tratamentos.

Com a crescente popularidade desta tecnologia, uma série de métodos foi desenvolvida para a detecção de genes diferencialmente expressos. Dentre os diversos métodos, apresentam destaque o edgeR (ROBINSON et al., 2010), DESeq (ANDERS e HUBER, 2010) e baySeq (HARDCASTLE e KELLY, 2010) que se baseiam na distribuição de probabilidade Binomial Negativa, e, as abordagens não-paramétricas, tais como NOIseq (TARAZONA et al, 2011) e o SAMseq (LI, 2013).

Diante das diversas metodologias disponíveis para análise de expressão diferencial, a comparação entre estas é importante e tem promovido o interesse entre os pesquisadores. Por exemplo, Kvam et al (2012) compararam os métodos estatísticos EdgeR, DESeq, baySeq, e o método com um modelo de Poisson de dois estágios (MPTA), em termos da sua capacidade para discriminar entre genes diferencialmente expressos e genes não diferencialmente expressos, de conjuntos de dados simulados baseados em diferentes distribuições de probabilidade ou de dados reais. Verificou-se que o MPTA não executa tão bem como os outros métodos quando o tamanho de amostra é pequeno. Robles et al. (2012) avaliaram o impacto do aumento da cobertura de sequenciamento para detectar genes diferencialmente expressos, contrastando com este as vantagens de aumentar o tamanho da amostra. Além disso, também foi verificado nesse trabalho o controle da taxa de falsos positivos desses métodos. Os resultados indicaram que o método baySeq tem menores taxas de falsos positivos. Soneson e Delorenzi (2013) conduziram um estudo onde foram comparados por meio de simulação de dados, onze métodos para análise de expressão diferencial com o intuito de verificar a influência do método de normalização na taxa de falsos-positivos (erro do tipo I) e o poder do teste. Como resultado, verificou-se que amostras pequenas, apresentaram problemas para todos os métodos avaliados.

Apesar dos estudos demonstrarem a importância da existência de repetições para a avaliação da expressão diferencial, na prática, muitos estudos ainda são realizados sem

repetições ou com poucas repetições. Com isso, o entendimento das melhores práticas continua a ser interessante e o campo está em contínuo desenvolvimento.

Diante do exposto, este trabalho tem por objetivo avaliar as metodologias DESeq (ANDERS e HUBER, 2010) e o baySeq (HARDCASTLE e KELLY, 2009) quanto ao número de repetições e verificar a coincidência dos resultados.

Tais metodologias foram escolhidas porque são as únicas que permitem análise sem repetição e caso fossem comparadas com as outras, não haveria a possibilidade de criar o cenário sem repetição, fugindo ao objetivo do trabalho.

2. Material e Métodos

Visando avaliar as técnicas de expressão diferencial, foram simulados conjuntos de observações contendo 1000 genes sob duas condições experimentais, por exemplo, tratado e não tratado. Foram considerados cenários denotados como C1, C2, C3 e C4 que correspondem a situações com cinco repetições, três repetições, duas repetições e sem repetição, respectivamente.

A simulação foi realizada por meio da função simulateReadCount do pacote TCC (SUN et al., 2013) do Bioconductor (GENTLEMAN et al., 2004), considerando que os 200 primeiros genes são diferencialmente expressos e 800 restantes não apresentam expressão diferencial. Além disso, considerou-se que os genes diferencialmente expressos diferem em 4 vezes dos não diferencialmente expressos. Dos 200 diferencialmente expressos simulados, 100 eram mais expressos em relação à referência (upregulated) e 100 menos expressos em relação à referência (downregulated). Para verificar a influência do número de repetições na detecção de genes diferencialmente expressos, avaliou-se a expressão diferencial por meio do DESeq (ANDERS e HUBER, 2010) e baySeq (HARDCASTLE e KELLY, 2010) considerando um FDR de 5%.

O procedimento foi repetido dez vezes, sendo utilizadas como resultados as médias das quantidades de genes diferencialmente expressos e não diferencialmente expressos detectados em cada cenário.

Posteriormente, com o intuito de avaliar concordância de genes diferencialmente expressos entre os quatro cenários avaliados e entre as metodologias foram construídos diagramas de Venn por meio do pacote Venndiagram do R (CHEN e BOUTROS, 2011).

Visando avaliar as metodologias quanto à sensibilidade (ou taxa de verdadeiros positivos), foram construídas as curvas ROC (Receiver Operating Characteristic) para cada método e

cada cenário. A área sob a curva ROC (AUC) foi usada como uma medida do desempenho global discriminativo frente aos genes não diferencialmente expressos.

3. Resultados e Discussão

3.1 DESeq

A redução no número de repetições leva ao decréscimo no percentual de acerto, quando se consideram apenas os genes diferencialmente expressos (Tabela 1). Especificamente, C1 (5 repetições) e C4 (sem repetições) apresentaram respectivamente o maior (57,8% - 116 genes diferencialmente expressos) e menor (1% - 2 genes diferencialmente expressos) percentual de acerto (Tabela 1). Esses resultados corroboram com o estudo realizado por Anders e Huber (2010), que avaliaram a eficiência da metodologia implementada no pacote DESeq para análises de dados de RNA-Seq sem repetições em células neurais e observaram que somente 11% dos genes foram considerados diferencialmente expressos, quando comparados com a análise realizada com duas repetições. Nesse mesmo trabalho, considerando dados de moscas sem repetições, os autores conseguiram identificar 75,09% dos genes considerados como diferencialmente expressos quando comparados ao conjunto de dados que trabalhavam com todas as repetições. Esses resultados indicam que a adequação da pressuposição feita para análise de dados sem repetições está diretamente ligada à variabilidade dos dados, evidenciando que a utilização do DESeq para experimentos sem repetições deve ser realizada com cautela visto que sua eficiência está ligada tanto ao número de repetições quanto a variabilidade das amostras (no caso sem repetição). Segundo Anders e Huber (2010), apesar do pacote DESeq permitir análises de experimentos biológicos sem repetição, em uma ou ambas as condições, não é possível tirar conclusões confiáveis dessas análises em termos estatísticos, visto que o uso apropriado de repetições é essencial para se interpretar um experimento biológico.

Tabela 1. Percentual de acerto para os genes diferencialmente expressos e taxa de falsos positivos

Cenário	% de acerto DE	Taxa de Falsos positivos	% de acerto DE Geral
C1	57,800%	0,360%	29,100%
C2	40,200%	0,840%	20,500%
C3	25,100%	1,940%	13,500%
C4	1,000%	6,390%	3,200%

Nota:C1: 5 repetições, C2: 3 repetições, C3: 2 repetições, C4: sem repetição

Quanto à taxa de falsos positivos, observou-se que no cenário C4 (sem repetição) a taxa é mais alta que para os demais canários. Na prática esse resultado indica que o pesquisador pode levar para análises posteriores genes que não são verdadeiramente diferencialmente expressos, gastando tempo e recursos com os mesmos.

Comparando os cenários entre si, em termos da interseção de genes diferencialmente expressos, observa-se que a probabilidade de encontrar os mesmos genes diferencialmente expressos foi maior entre os cenários C1 e C2 (5 repetições e 3 repetições, respectivamente) (Figura 1A), com 24% de concordância. Entre os cenários C1 e C3, o percentual de concordância foi de 17, 5% e entre os cenários C2 e C3 foi de 21,5%. Não houve concordância com o cenário C4 (sem repetição).

Considerando-se os genes não DE observa-se que em geral os genes não DE quando se consideram poucas repetições são os mesmos indicados quando a análise é realizada com um número maior de repetições (Figura 1B).

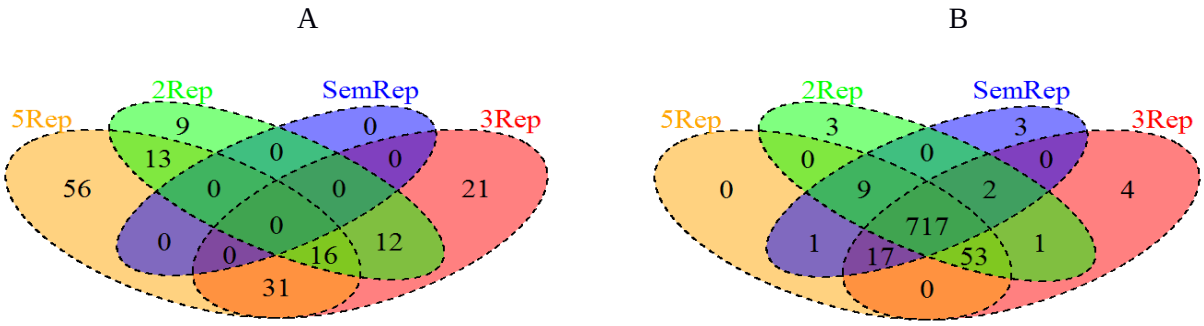

Figura 1. Diagrama de Venn para genes diferencialmente expressos (A) e Não diferencialmente expressos (B)

3.2 baySeq

Os resultados obtidos pelo baySeq mostram que a redução no número de repetições promove a diminuição no percentual de acertos, considerando-se os genes diferencialmente expressos. Os cenários C1 (5 repetições) e C4 (sem repetições) apresentaram respectivamente, o maior (59% - 118 acertos) e menor (0,00% - nenhum acerto) percentual de acerto (Tabela 2). Esses resultados são semelhantes aqueles obtidos pelo DESeq, evidenciando a importância de repetições biológicas em experimentos dessa natureza.

Tabela 2. Percentual de acerto para os genes diferencialmente expressos e taxa de falsos positivos

Cenário	% de acerto DE	Taxa de Falsos positivos	% de acerto DE Geral
C1	59,000%	0,660%	29,80%
C2	40,200%	0,900%	20,500%
C3	23,800%	2,080%	12,900%
C4	0,000%	6,390%	3,200%

Nota: C1: 5 repetições, C2: 3 repetições, C3: 2 repetições, C4: sem repetição.

À medida que decresce o número de repetições, há um aumento na taxa de falsos positivos, sendo que no cenário C4 essa taxa é ainda maior. Esses resultados mostram que mesmo modificando o método de análise, quando há decréscimo no número de repetições a taxa de acerto de genes diferencialmente expressos cai consideravelmente.

Em se tratando do grau de concordância entre de genes DE, os percentuais entre os cenários foram de 21,7% entre C1 e C2; 16,8% entre C1 e C3 e 17,9% entre C2 e C3 (Figura 2A). Para os genes não diferencialmente expressos, os resultados indicam também que quando se consideram poucas repetições, esses são os mesmos indicados quando a análise é realizada com um número maior de repetições (Figura 2B).

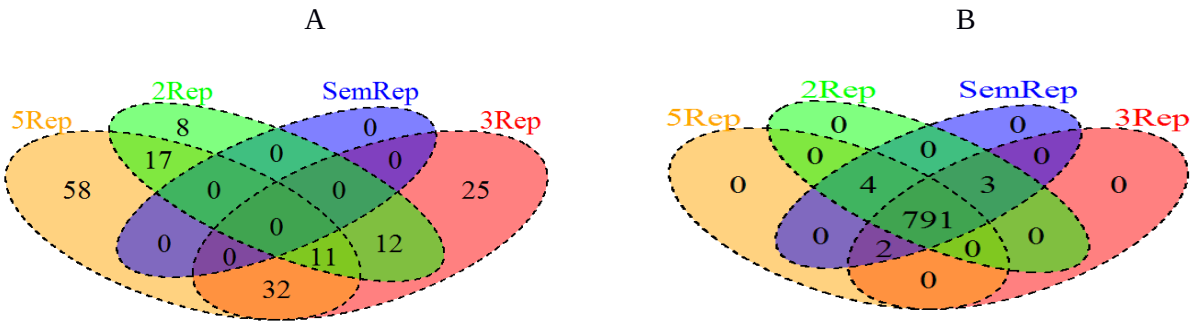

Figura 2. Diagramas de Venn para genes diferencialmente expressos (A) e Não diferencialmente expressos (B).

3.3 DESeq x baySeq

O baySeq apresentou mais genes como diferencialmente expressos do que o DESeq, sendo que as taxas de falsos positivos foram bastante aproximadas (Tabelas 1 e 2). Tais resultados corroboram com o estudo comparativo realizado por Seyednasrollah et al. (2013), entre os pacotes edgeR, DESeq, baySeq, NOIseq, SAMseq, limma, Cuffdiff 2 e EBSeq e

concluíram que o DESeq é um pouco mais conservador, ou seja, detecta menos diferencialmente expressos.

A maioria dos genes considerados como diferencialmente expressos pelo DESeq também o foram pelo baySeq (Figura 3). Outros trabalhos comparativos já testaram tais metodologias em diversas situações e concluíram que nenhum dos métodos é ótimo em todas as circunstâncias e que a escolha do método depende das condições experimentais (RAPAPORT, et al., 2013; SONESON e DELORENZI, 2013). O fato de o DESeq ter detectado menor número de genes diferencialmente expressos está consistente com o fato de ele ser bastante conservador, corroborando com o trabalho realizado por Soneson e Delorenzi (2013).

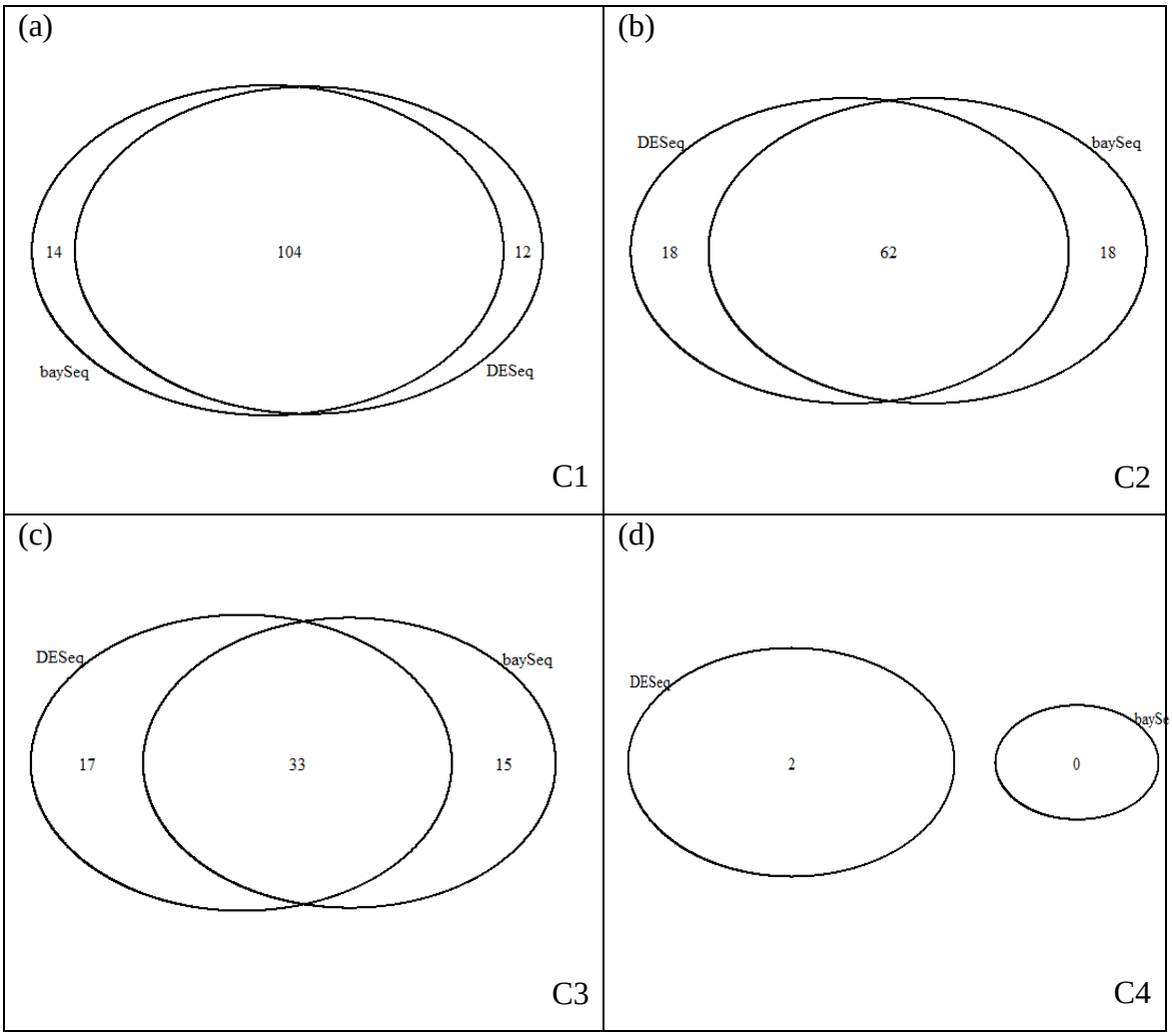

Figura 3 - Sobreposição dos genes diferencialmente expressos entre os métodos, por número de repetições, com FDR menor que 5%.

3.4 Curvas ROC

Em todos os cenários o método baySeq apresentou um melhor desempenho do que o DESeq para classificar genes verdadeiramente diferencialmente expressos, sendo que no cenário C4 o desempenho do baySeq (AUC=0,718) foi significativamente maior do que o do DESeq (AUC=0,502), indicando que para experimentos sem repetição o baySeq seria o método mais indicado, com uma taxa menor de falsos positivos. Esses resultados devem ser vistos com certa cautela, tendo em vista que na comparação feita por Seyednasrollah et al. (2013), apesar de concluir que o DESeq é mais conservador, os autores afirmam que o pacote baySeq apresentou uma alta variabilidade.

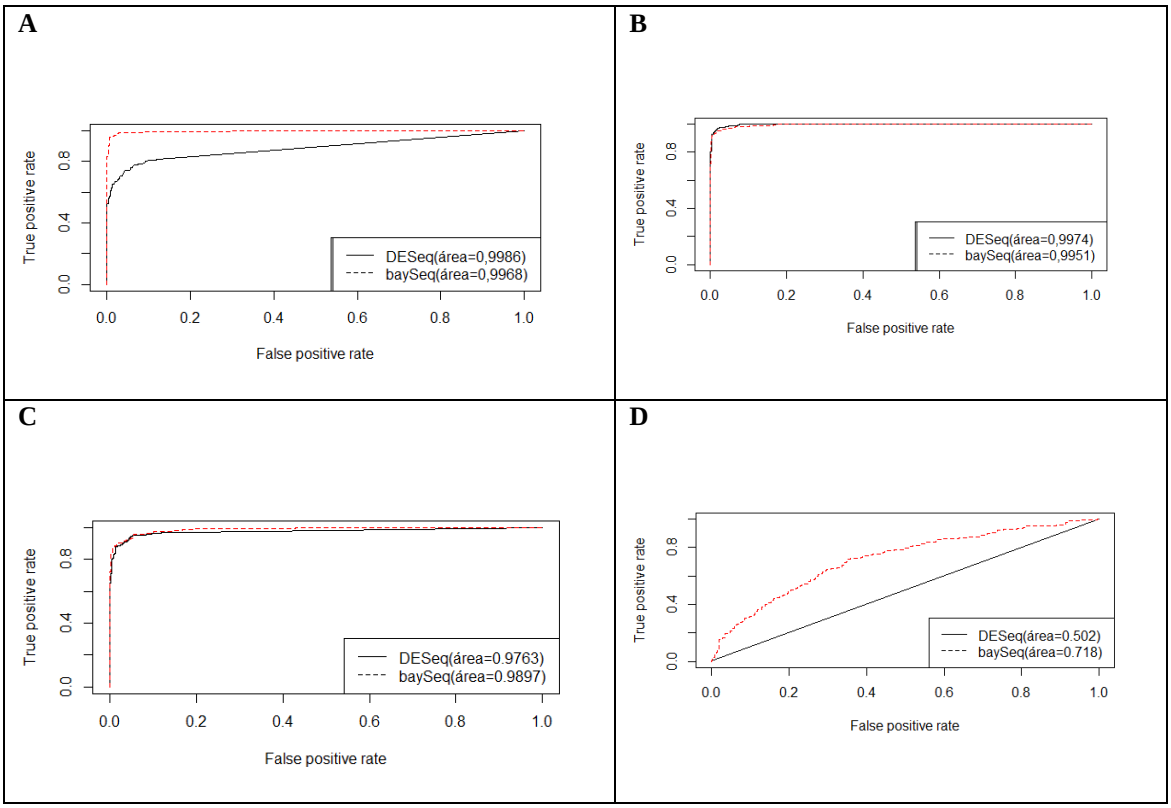

Figura 4 – Curvas ROC para os cenários C1, C2, C3, e C4 (A, B, C e D)

4. CONCLUSÕES

Neste trabalho, foram avaliados e comparados os métodos, baySeq e DESeq, para análise de expressão diferencial de experimentos de RNA-Seq.

Sob as condições analisadas, o método abordado pelo baySeq foi o que apresentou um desempenho um pouco melhor nos cenários onde havia repetições, e no cenário sem repetições, apresentou uma maior sensibilidade, ou seja, maior taxa de verdadeiros positivos e menor taxa de falsos positivos.

Referências Bibliográficas

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

BENJAMINI, Y.; HOCHBERGH,Y. **Controlling the false discovery rate - a practical and powerful approach to multiple testing**. JRStat Soc BMethodol 1995;57:289–300.

CHEN H., BOUTROS P.C. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. **BMC Bioinformatics** 12: 35

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**, 5:R80. 2004.

HARDCASTLE, T. J.; KELLY, K. A. Bayseq: Empirical bayseian methods for identifying differential expression in sequence count data. **BMC Bioinformatcs**, 11 R: 106, 2010.

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.

LI, J.; TIBSHIRANI, R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. **StatMethodsMed Res** 2013;22:519–523.

LOVE, M., ANDERS, S., and HUBER, W. (2013). **Differential analysis of count data - the DESeq2 package**

http://www.bioconductor.org/packages/2.13/bioc/vignettes/deseq2/inst/doc/deseq2.pdf

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.

NOOKAEW, I.; PAPINI, M.; PORNPUTTAPONG, N**.** A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. **Nucleic Acids Res** 2012;40:10084–10097

RAPAPORT, F.; KHANIN, R.; LIANG, Y.; PIRUN, M.; KREK, A.; ZUMBO, P.; MASON, C. E.; SOCCI, N. D.; BETEL, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-Seq data. **Genome biology**, 14(9), R95.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. **EdgeR: a bioconductor package for differential expression analysis of digital gene expression data.** Bioinformatics, 26:139–140. 2010.

SEYEDNASROLLAH, F.; LAIHO, A.; ELO, L.L**. Comparison of software packages for detecting differential expression in RNA-Seq studies**. Brief Bioinform 2013.

SMYTH, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. **Stat Appl GenetMol Biol** 2004;3:Article 3.

SONESON, C.; DELORENZI, M. A comparison of statistical methods for detecting diferentially expressed genes from RNA-Seq data. **BMC Bioinformatics**, 14:91, 2013.

SUN, J.; NISHIYAMA, T.; SHIMIZU, K.; KADOTA, K. (2013) TCC: An R Package for Comparing Tag Count Data with Robust Normalization Strategies. **BMC Bioinformatics**, 14, 219. http://dx.doi.org/10.1186/1471-2105-14-219.

TARAZONA, S.; GARCÍA, A. F.; DOPAZO, J.; FERRER, A.; CONESA, A. Differential expression in RNA-Seq: a matter of depth. **Genome Res**, 21:2213–2223. 2011.

APÊNDICE A – Rotinas computacionais implementadas

As principais rotinas computacionais dos métodos descritos neste trabalho foram implementadas no software livre R (R Development Core Team, 2012) e estão descritas a seguir.

1. **Criação do banco de dados:**

```
ngenes=1000
pgde=0.2
fc=c(4,4)
repeticoes5=c(5,5)
repeticoes3=c(3,3)
repeticoes2=c(2,2)
repeticoes1=c(1,1)
```

Deseq

5 repetições

```
tcc5<-simulateReadCounts(Ngene = ngenes, PDEG = pgde,DEG.assign = c(0.5, 0.5),DEG.foldchange = fc,replicates = repeticoes5)
cont5<-tcc5$count
cont5 <- data.frame(cont5)
delineamento5 = data.frame(row.names = colnames(cont5),condition5 = c("untreated","untreated","untreated","untreated","untreated", "treated","treated","treated","treated","treated"))
condition5= delineamento5$condition5
cds5<-newCountDataSet(cont5, condition5)
cds5= estimateSizeFactors(cds5)
sizeFactors(cds5)
cds5 = estimateDispersions(cds5)
head(fData(cds5))
cds5 = estimateDispersions(cds5)
res5 = nbinomTest(cds5, "untreated", "treated")
acerto_DE_5_DESeq=sum(res5[1:200,8]<0.05,na.rm = T)
acerto_DE_5_DESeq
acerto_NDE_5_DESeq=sum(res5[201:1000,8]>0.05,na.rm = T)
```

```
acerto_NDE_5_DESeq
perc_DE_5_DESeq= acerto_DE_5_DESeq/200
perc_DE_5_DESeq
perc_NDE_5_DESeq= acerto_NDE_5_DESeq/800
perc_NDE_5_DESeq
tfp5_DESeq=1-perc_NDE_5_DESeq
tfp5_DESeq
```

Obs.: Para os cenários com 3 e 2 repetições, o procedimento é o mesmo que para 5 repetições, bastando substituir o 5 por 3 e 2.

#sem repetição

```
tcc1 <- simulateReadCounts(Ngene = ngenes, PDEG = pgde,DEG.assign = c(0.5, 0.5),DEG.foldchange = c(4,4),replicates = repeticoes1)
cont1<-tcc1$count
cont1 <- data.frame(cont1)
delineamento1 = data.frame(row.names = colnames(cont1),condition = c("untreated","treated"))
condition = delineamento1$condition
cds1<-newCountDataSet(cont1, condition)
cds1 = estimateSizeFactors(cds1)
sizeFactors(cds1)
cds1 = estimateDispersions(cds1, method = "blind", sharingMode = "fit-only")
head(fData(cds1))
res1 = nbinomTest(cds1, "untreated", "treated")
head(res1)
acerto_DE_1_DESeq=sum(res1[1:200,8]<0.05,na.rm = T)
acerto_DE_1_DESeq
acerto_NDE_1_DESeq=sum(res1[201:1000,8]>0.05,na.rm = T)
acerto_NDE_1_DESeq
```

```
perc_DE_1_DESeq= acerto_DE_1_DESeq/200
perc_DE_1_DESeq
perc_NDE_1_DESeq= acerto_NDE_1_DESeq/800
perc_NDE_1_DESeq
tfp1_DESeq=1-perc_NDE_1_DESeq
tfp1_DESeq
```

Bayseq

5 repetições

```
dados5<-as.matrix(cont5)
dados5<-calcNormFactors(dados5)
rep5 <- c("simA", "simA", "simA", "simA", "simA", "simB",
"simB", "simB", "simB", "simB")
groups5 <- list(NDE = c(1,1,1,1,1,1,1,1,1,1), DE =
c(1,1,1,1,1,2,2,2,2,2))
CD5 <- new("countData", data = dados5, replicates = rep5,
groups = groups5)
libsizes(CD5) <- getLibsizes(CD5)
CDPriors5 <- getPriors.NB(CD5, cl = NULL)
CDPost5 <- getLikelihoods.NB(CDPriors5, cl = NULL)
a5=topCounts(CDPost5, group = "DE",number=1000)
a5=by(a5, a5[,"rowID"],function(x) x)
x5<-do.call("rbind",as.list(a5))
head(x5)
fdr5<-(x5[,14])
head(fdr5)
acerto_DE_5_baySeq=sum(x5[1:200,14]<0.05)
acerto_DE_5_baySeq
acerto_NDE_5_baySeq=sum(x5[201:1000,14]>0.05)
acerto_NDE_5_baySeq
perc_DE_5_baySeq= acerto_DE_5_baySeq/200
perc_DE_5_baySeq
perc_NDE_5_baySeq= acerto_NDE_5_baySeq/800
perc_NDE_5_baySeq
```

```
tfp5_baySeq=1-perc_NDE_5_baySeq
tfp5_baySeq
```

Obs: A simulação dos dados para 3 repetições e 2 repetições foram realizadas da mesma forma acima, somente modificando o número de repetições

Sem repetição

```
dados1<-as.matrix(cont1)
dados1<-calcNormFactors(dados1)
rep1<- c("simA","simB")
groups1 <- list(NDE = c(1,1), DE = c(1,2))
CD1 <- new("countData", data = dados1, replicates = rep1, groups = groups1)
libsizes(CD1) <- getLibsizes(CD1)
CDPriors1 <- getPriors.NB(CD1, cl = NULL)
CDPost1 <- getLikelihoods.NB(CDPriors1, cl = NULL)
a1=topCounts(CDPost1, group = "DE",number=1000)
a1=by(a1, a1[,"rowID"],function(x) x)
x1<-do.call("rbind",as.list(a1))
head(x1)
fdr1<-(x1[,6])
head(fdr1)
acerto_DE_1_baySeq=sum(x1[1:200,6]<0.05)
acerto_DE_1_baySeq
acerto_NDE_1_baySeq=sum(x1[201:1000,6]>0.05)
acerto_NDE_1_baySeq
perc_DE_1_baySeq= acerto_DE_1_baySeq/200
perc_DE_1_baySeq
perc_NDE_1_baySeq= acerto_NDE_1_baySeq/800
perc_NDE_1_baySeq
tfp1_baySeq=1-perc_NDE_1_baySeq
tfp1_baySeq
```

```
#Encontrar a interseção
#genes DE
#5 repetições
aa51=(DESeq51[1:200,7]<0.05)
bb51=(Bayseq51[1:200,14]<0.05)
ee51=cbind(aa51,bb51)
eee51<- apply(ee51,1,sum)
acertos51<-apply(ee51,2,sum)
acertos51
int51=sum(eee51==2,na.rm=TRUE)
int51
#3 repetições
aa31=(DESeq31[1:200,8]<0.05)
bb31=(Bayseq31[1:200,10]<0.05)
ee31=cbind(aa31,bb31)
ee311<-na.exclude(ee31)
eee311<- apply(ee311,1,sum)
acertos31<-apply(ee311,2,sum)
acertos31
int31=sum(eee31==2,na.rm=TRUE)
int31
#2 repetições
aa21=(DESeq21[1:200,8]<0.05)
bb21=(Bayseq21[1:200,8]<0.05)
ee21=cbind(aa21,bb21)
eee21<- apply(ee21,1,sum)
acertos21<-apply(ee21,2,sum)
acertos21
int21=sum(eee21==2,na.rm=TRUE)
int21
#sem repetição
aa11=(DESeq11[1:200,7]<0.05)
bb11=(Bayseq11[1:200,6]<0.05)
ee11=cbind(aa11,bb11)
```

```
ee111=na.exclude(ee11)
eee111<- apply(ee111,1,sum)
acertos11<-apply(ee111,2,sum)
acertos11
int11=sum(eee111==2,na.rm=TRUE)
int11
#Encontrar a interseção
#genes NDE
#5 repetições
aa51=(DESeq51[201:1000,8]>0.05)
bb51=(Bayseq51[201:1000,14]>0.05)
ee51=cbind(aa51,bb51)
ee511=na.exclude(ee51)
eee511<- apply(ee511,1,sum)
acertos51<-apply(ee511,2,sum)
acertos51
int51=sum(eee511==2,na.rm=TRUE)
int51
#3 repetições
aa31=(DESeq31[201:1000,8]>0.05)
bb31=(Bayseq31[201:1000,10]>0.05)
ee31=cbind(aa31,bb31)
ee3111<-na.exclude(ee31)
eee311<- apply(ee311,1,sum)
acertos31<-apply(ee311,2,sum)
acertos31
int31=sum(eee311==2,na.rm=TRUE)
int31
#2 repetições
aa21=(DESeq21[201:1000,8]>0.05)
bb21=(Bayseq21[201:1000,8]>0.05)
ee21=cbind(aa21,bb21)
ee211=na.exclude(ee21)
eee211<- apply(ee211,1,sum)
```

```
acertos21<-apply(ee211,2,sum)
acertos21
int21=sum(eee211==2,na.rm=TRUE)
int21
#sem repetição
aa11=(DESeq11[201:1000,7]>0.05)
bb11=(Bayseq11[201:1000,6]>0.05)
ee11=cbind(aa11,bb11)
ee111=na.exclude(ee11)
eee111<- apply(ee111,1,sum)
acertos11<-apply(ee111,2,sum)
acertos11
int11=sum(eee111==2,na.rm=TRUE)
int11
```