PÂMELA TAMIRIS CALDAS SERRA DE SOUZA

COMPARAÇÃO DE METODOLOGIAS PARA IDENTIFICAÇÃO DE GENES DIFERENCIALMENTE EXPRESSOS EM EXPERIMENTOS DE RNA-Seq DE SUÍNOS

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

VIÇOSA
MINAS GERAIS - BRASIL
2015

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

T
S729c
2015

Souza, Pâmela Tamiris Caldas Serra de, 1989-
Comparação de metodologias para identificação de genes diferencialmente expressos em experimentos de RNA-Seq de suínos / Pâmela Tamiris Caldas Serra de Souza. – Viçosa, MG, 2015.
vii, 37f. : il. ; 29 cm.

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

1. Estatísticas. 2. Biometria. 3. Metodologias - Análise. 4. Bioinformática. 5. Suínos. 6. Sequenciamento de nucleotídeos. 7. Regulação da 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

PÂMELA TAMIRIS CALDAS SERRA DE SOUZA

COMPARAÇÃO DE METODOLOGIAS PARA IDENTIFICAÇÃO DE GENES DIFERENCIALMENTE EXPRESSOS EM EXPERIMENTOS DE RNA-Seq DE SUÍNOS

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

APROVADA: 08 de abril de 2015.

_____ _____

Ana Carolina Campana Nascimento Talles Eduardo Ferreira Maciel

_____

Moysés Nascimento
(Orientador)

À Deus e a minha mãe Dinete,
por toda dedicação para
com o meu desenvolvimento
pessoal e profissional.

AGRADECIMENTOS

À 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.

Ao Departamento de Zootecnia da Universidade Federal de Viçosa, pela concessão dos dados utilizados na pesquisa.

À minha mãe, Dinete, pelo amor incondicional, ensinamentos e confiança, sem a qual eu jamais chegaria aqui.

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

Aos meus amigos Nayara, Wagner, Gabi, Laís, Vanessa, Esteferson e distantes, pela amizade, companheirismo, carinho, incentivo e pelos bons momentos que passamos juntos.

Aos orientadores Moysés Nascimento e Fabyano Fonseca e Silva, pelos ensinamentos, confiança, dedicação e por contribuir para o meu crescimento profissional, além de serem grandes exemplos a ser seguido.

Aos membros da banca examinadora, Ana Carolina Campana Nascimento e Talles Eduardo Ferreira Maciel, 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 CAPES, pela concessão da bolsa de estudos.

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

BIOGRAFIA

PÂMELA TAMIRIS CALDAS SERRA DE SOUZA, filha de Dinete Caldas Serra e Gilvan de Souza, nasceu em Monte Dourado, Pará, em 21 de outubro de 1989.

Em março de 2008, ingressou no curso de Bacharelado em Estatística na Universidade Federal do Pará, Pará – PA, graduando-se em dezembro de 2011.

Em fevereiro de 2012, 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 08 de abril de 2015.

SUMÁRIO

RESUMO

SOUZA, Pâmela Tamiris Caldas Serra, M.Sc., Universidade Federal de Viçosa, abril de 2015. **Comparação de Metodologias para Identificação de Genes Diferencialmente Expressos em Experimentos de RNA-Seq de Suínos**. Orientador: Moysés Nascimento. Coorientadores: Fabyano Fonseca e Silva, Fernando Luiz Pereira de Oliveira e Wagner Antonio Arbex.

Um dos principais desafios da biologia molecular é medir e avaliar os perfis de expressão gênica em diferentes condições com o objetivo de entender os mecanismos de transformação molecular. Para tanto, o método RNA-Seq usa o transcriptoma obtido a partir de tecnologias de sequenciamentos de nova geração (NGS), as quais são utilizadas para converter RNA em uma biblioteca de fragmentos de cDNA, e, assim, produzir milhões reads. Após a mensuração dos níveis de expressão dos genes, por meio de técnicas de mapeamento, surge a necessidade de verificar hipóteses a respeito da existência de expressão diferencial (ED) entre as condições avaliadas. Assim, faz-se necessária à descoberta e o aprimoramento de metodologias estatísticas para aperfeiçoar as análises de dados gerados em plataformas de sequenciamento de genomas. O objetivo geral desse estudo consistiu em avaliar o comportamento de três metodologias (DEGSeq, bayseq e DESeq) para verificação da expressão diferencial em longissimus dorsi (LD) do músculo de suínos da raça Piau e Comercial, em 21e 90 dias depois do coito, por meio de dados provenientes de RNA-Seq, em cenários sem repetição . De acordo com os resultados gerados nas análises e sob as condições utilizadas no desenvolver do experimento concluiu-se que, na comparação dos métodos bayseq com DEGSeq e baySeq com DESeq, respectivamente, observou-se, a partir da relação do nível de expressão (fold-change) entre as duas raças suínas (comercial e piau), que os métodos apresentaram desempenho diferentes entre si, pois apresentaram um nível de expressão desigual em ambos os métodos. No entanto, na comparação entre os métodos DESeq e DEGSeq, houve um desempenho comparável, deste modo, houve concordância entre os métodos. Como um todo, a maioria dos genes DE identificados, se deu na fase pós-natal tardia, ou seja, 90 dpc. Além disso, a maioria deles foram down na fase pré-natal inicial (21 dpc) e foram up na fase pré-natal tardia (90 dpc) relacionando as raças, comercial e piau e comparando os métodos.

ABSTRACT

SOUZA, Pâmela Tamiris Caldas Serra, M.Sc., Universidade Federal de Viçosa, april de 2015. **Comparison Methodologies for Identification Genes Differentially Expressed in RNA-Seq Experiments of Pigs**. Adviser: Moysés Nascimento. Co-Advisers: Fabyano Fonseca e Silva, Fernando Luiz Pereira de Oliveira and Wagner Antonio Arbex.

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 RNA-Seq method uses transcriptome from young generation sequencing technologies (NGS), used to convert RNA into a cDNA fragment library, and thus produce millions reads. After the measurement of levels of gene expression, the need arises to test hypotheses about the existence of differential expression (DE) between the evaluated conditions. Thus, it is necessary to the discovery and improvement of efficient statistical methods to improve data analysis generated by genome sequencing platforms. The overall objective of this study was to evaluate the behavior of three methodologies (DEGSeq, bayseq and DESeq) to verify the differential expression in longissimus dorsi (LD) muscle of the pig Piau and Commercial race in 21 and 90 days after intercourse, by using data from RNA Seq in scenarios without repetition. According to the results generated by the analysis and under the conditions used to develop the experiment it was concluded that, in comparison with the methods bayseq DEGSeq and baySeq with DESeq respectively, was observed from the relation of expression level (fold-change) between the two pig breeds (commercial and piau), the methods showed different performance between them, they showed an uneven level of expression in both methods. However, when comparing the DESeq and DEGSeq methods, there was a comparable performance thus there was agreement between the methods. As a whole, the majority of the identified genes occurred in the late post-natal period, namely 90 dpc. Moreover, most of them were down in early postnatal stage (21 dpc) were up in late postnatal period (90 dpc) relating races, commercial and piau and comparing the methods.

1. INTRODUÇÃO

O transcriptoma é o conjunto de transcritos e suas quantidades em um estágio específico do desenvolvimento ou condição fisiológica; sendo assim, o reflexo direto da expressão gênica. O estudo do transcriptoma permite: catalogar transcritos, identificar e determinar a estrutura dos genes e quantificar mudanças nos níveis de expressão gênica sob diferentes condições (WANG et al., 2009).

Em meados da década de noventa, a técnica de microarray era utilizada como principal ferramenta para mensuração de transcritos e posterior entendimento dos perfis de expressão de grande número de genes, em diferentes condições (SCHENA et al., 1995). Embora tal técnica permita verificar de forma rápida e simultânea os níveis de expressão de milhares de genes, a mesma apresenta algumas limitações, como por exemplo: necessidade de conhecimento prévio das sondas utilizadas para hibridização, mensuração relativa da expressão gênica (uma vez que se mede a intensidade de luz emitida), susceptibilidade à hibridização cruzada, baixa reprodutibilidade de resultados entre laboratórios e ineficiência no estudo de genes raros e de isoformas (ESTEVES, 2007; FERREIRA FILHO, 2009; NEVES, 2010).

A técnica de microarray vem sendo substituída pelo sequenciamento de RNA em larga escala (RNA-Seq). Tal técnica não necessita do conhecimento prévio das sequências dos genes do organismo em estudo e utiliza diretamente a contagem das sequências dos RNAs sequenciados como medida de expressão de cada gene; apresentando assim maior reprodutibilidade de resultados.

De maneira sucinta, a metodologia de RNA-Seq se caracteriza por converter RNA em uma biblioteca de fragmentos de cDNA, de forma que cada molécula pode ser sequenciada gerando pequenas sequências (reads) com tamanho variando entre 21 e 500 bp (WANG et al., 2009).

Após a mensuração dos níveis de expressão dos genes, surge a necessidade de verificar hipóteses a respeito da existência de expressão diferencial (ED) entre as condições avaliadas. Para tanto, diversos métodos baseados na suposição da normalidade dos dados (intensidade de luz emitida) foram propostos quando os níveis de expressão são provenientes da técnica de microarray (YANG et al., 2001 e 2002). Contudo, em análises de expressão gênica por meio de RNA-Seq, a medida de expressão é uma variável discreta e se refere ao número de reads alinhados a determinado gene. Assim, visto que a transformação dos dados de contagem não é bem aproximada por distribuições contínuas, modelos estatísticos apropriados para dados de contagem foram propostos para extrair o máximo de informações a partir de dados provenientes do RNA-Seq.

Inicialmente, as metodologias de expressão diferencial modelavam os dados de contagem, provenientes do RNA-Seq, por meio da distribuição de Poisson (MARIONI, 2008). Entretanto, a variabilidade não é bem modelada por meio dessa distribuição uma vez que a mesma possui um único parâmetro, o qual é exclusivamente determinado pela sua média, tornando assim a variância igual à média (LANGMEAD et al., 2010). Este fato faz com que as análises baseadas na distribuição de Poisson fiquem propensas a altas taxas de falsos positivos (ROBINSON et al., 2010).

Visando uma modelagem mais adequada, reduzindo o número de falsos positivos, a literatura apresenta diversos métodos baseados na distribuição binomial negativa, que é uma extensão da distribuição de Poisson, porém requer um parâmetro de variância, diferente da média, a ser estimado (BULLARD, 2010).

Dentre as diversas metodologias para o estudo da expressão diferencial, as baseadas em distribuições binomial negativa (ROBINSON e SMYTH, 2008) apresentam grande destaque e são amplamente utilizadas. Metodologias essas, que diferem quanto aos conceitos adotados e aos princípios estatísticos empregados. Como exemplo, pode-se citar as metodologias DEGSeq (WANG et al., 2010), baySeq (HARDCASTLE e KELLY, 2009) e DESeq (ANDERS e HUBER, 2010).

Diante da grande quantidade de metodologias e da falta de um consenso sobre qual a melhor para avaliar ED, alguns trabalhos têm sido desenvolvidos a fim de compará-las em relação à capacidade de detectar genes diferencialmente expressos. Como por exemplo, o estudo desenvolvido por Sonenson e Delorenzi (2013) em que foram comparados onze métodos para a análise da expressão diferencial, baseados na distribuição binomial negativa e Poisson, em dados simulados e reais de RNA-seq, visando avaliar o impacto que a normalização causa nos resultados da análise de expressão diferencial. Como resultado, verificou-se que para amostras muito pequenas os resultados obtidos devem ser interpretados com cautela. Para tamanhos maiores de amostra, os métodos que utilizam duas abordagens de transformação (a transformação variância estabilizada pelo método DESeq e a transformação voom do pacote limma) (SMYTH, 2004) e o método SAMseq não paramétrico (LI e TIBSHIRANI, 2011) tiveram um melhor desempenho sobas diversas condições avaliadas, sendo relativamente pouco afetados por valores discrepantes.

Kvam et al. (2012) compararam o edger (ROBINSON et al., 2010), DESeq (ANDES e HUBER, 2010) e baySeq (HARDCASTLE e KELLY, 2010) e um método com base em um modelo Poisson de dois estágios MPTA (AUER e DOERGE, 2011) os quais se baseiam na distribuição de poisson e binomial negativa em dados simulados e reais. Comparou-se a capacidade destes métodos para detectar genes DE em termos de ranking de significância de

genes e controle da taxa de falsa descoberta (FDR). Verificando que o método baySeq apresentou um melhor desempenho em termos de ranking de genes como sendo declarados DE, especialmente taxas de falsos positivos, que é de importância mais prática. Edger e DESeq obtiveram um desempenho semelhante e perto de baySeq. Os resultados de MPTA são mais variáveis e frequentemente o mais pobre quando o número de repetições é pequeno.

Estudos de ED por meio do RNA-Seq vêm sendo realizados em diversas áreas da ciência, como por exemplo, em medicina (BAILÃO, 2008), agronomia (SILVA, 2014) e zootecnia (FERRAZ 2009). Especificamente, no melhoramento animal, as análises de RNA-Seq podem ser utilizadas em estudos nos quais se objetivam caracterizar o desenvolvimento muscular de suínos, principalmente, por meio da identificação de genes diferencialmente expressos controladores da miogênese (processo de formação dos músculos) submetidos a diferentes raças.

Segundo BAXTER et al.(2008) o estudo do desenvolvimento muscular na fase pré-natal é fundamental para o entendimento de características de importância econômica, tais como sobrevivência e peso ao nascer, isso porque este é considerado um período de desenvolvimento máximo dos tecidos e órgãos, sendo influenciado tanto por fatores genéticos quanto ambientais (GREENWOOD et al., 2010).Visando tal entendimento, Solero et al. (2011), por meio de experimentos de microarray, identificaram genesque apresentam ED em músculo longissimus dorsi (LD) de suínos em 40 e 70 dias de gestação em suínos da raça Piau e Yorkshire-Landrace.

Apesar de útil, como já discutido, a técnica de microarray apresenta algumas limitações quando comparada ao RNA-Seq. Assim, o estudo do desenvolvimento muscular na fase pré-natal em suínos por meio de dados provenientes da técnica de RNA-Seq torna-se importante, visando à obtenção de um melhor resultado, verificando o comportamento das metodologias em termos do número de genes que apresentam expressão diferencial e ainda, se os genes identificados são mesmos para as diferentes metodologias utilizadas.

Diante do exposto, o objetivo geral deste trabalho é avaliar o comportamento de três metodologias (DEGSeq, bayseq e DESeq) para verificação da expressão diferencial em músculo longissimus dorsi (LD) de suínos da raça Piau e Comercial,em 21 e 90 dias depois do coito (dpc), por meio de dados provenientes de RNA-Seq.

CAPÍTULO 1

2. REVISÃO DE LITERATURA

2.1. RNA-Seq a partir das tecnologias de sequenciamento de nova geração (NGS)

O sequenciamento de RNA está sendo cada vez mais utilizado, fazendo com que o método seja inovador em pesquisas de transcriptomas, pois além de proporcionar uma maior sensibilidade que técnicas anteriores, como por exemplos os microarrays, o mesmo não necessita de uma lista pré-definida dos genes que se deseja detectar e não se limita apenas a avaliação de genes para os quais existam sondas (BULLARD et al., 2010, MARIONI et al., 2008). A princípio, qualquer transcrito que estiver sendo expresso pode ser detectado, 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, se o experimento tiver cobertura suficiente.

Essas tecnologias permitem ainda o estudo de vários fenômenos biológicos, incluindo polimorfismo de nucleotídeo único[1] (SNP), eventos epigenéticos[2], splicing[3]alternativo e o estudo de interações proteína-DNA [4](WANG et al., 2009 apud GONÇALVES, 2013).

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. Essas sequências podem conter, além do DNA de interesse, segmentos de clonagem que necessitam ser identificados e removidos antes de qualquer estudo, para se evitar interpretações errôneas (FALEIRO et al., 2011). De acordo com Amaral (2015), 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:

- ✓ Trimagens de reads duplicadas.
- ✓ Trimagem de adaptadores.

[1] Variação na sequência de DNA que afeta somente uma base nitrogenada na sequência dos genomas.
[2]São alterações com caráter reversível que não provocam modificações na sequência de DNA, mas sim no fenótipo, exercendo assim, influência nos mecanismos de expressão gênica.
[3]É um processo que remove os íntrons e junta os éxons depois da transcrição do RNA.
[4] Interação de DNA com inúmeras proteínas que desempenham suas funções em conjunto com o DNA.

- ✓ Exclusão de sequências super representadas.
- ✓ Trimagem de extremidades de reads com valor de Phred menor que 20.
- ✓ Trimagem de reads contendo mais que uma base ambígua, representada no conjunto de dados pela letra "N".
- ✓ Trimagem de reads com valor de Phred médio menor que 20.
- ✓ Trimagem das reads menores que 50 pares de bases.

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 usar os dados de RNA-seq tendo por objetivo comparar a expressão gênica sob determinadas condições, é necessário transformar milhões de reads em uma quantificação de expressão. O primeiro passo neste processo é o mapeamento de trechos genômicos sequenciados a partir dos fragmentos de interesse capturados nos sequenciadores de nova geração, denominados reads, com o intuito de encontrar o local único onde cada read melhor se alinha à referência (CASTAN, 2014).

2.2.3. Normalização

A normalização é uma etapa essencial na análise da expressão gênica diferencial visto às diferentes quantidades do número de reads entre as diferentes condições a serem comparadas, mais especificamente, dos diferentes tamanhos da biblioteca entre e dentre as amostras (Dillies et al., 2012).

Em particular, maiores coberturas resultam numa maior contagem para toda a amostra, influenciando no número de reads mapeadas a cada transcrito. Podendo provocar erros nas análises, devido à ocorrência de alterações no ranking de genes diferencialmente expressos, causados pelo comprimento do transcrito (OSHLACK E WAKEFIELD, 2009).

A normalização tem por objetivo permitir que os níveis de expressão gênica entre e dentre as amostras sejam comparáveis (SULTAN et al., 2008; MORTAVAZI et al., 2008; MARIONI et al., 2008).

Durante os últimos três anos, têm surgido várias abordagens de normalização para tratar os dados de RNA-Seq que diferem no tipo de ajuste e na estratégia estatística adotada. 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).

2.2.4. Análise da expressão gênica diferencial

Após o tratamento dos dados, mapeamento e obtenção da tabela de contagem, o interesse recai em testar hipóteses a respeito da existência de expressão diferencial (ED) entre as diferentes condições avaliadas. Para tanto, diante da natureza discreta da variável (número de reads) os métodos propostos para tal avaliação se baseiam em distribuições discretas de probabilidade, tais como Binomial, Poisson e Binomial Negativa.

Inicialmente, as metodologias modelavam os dados por meio da distribuição de Poisson (MARIONI, 2008). Entretanto, visto que a mesma possui um único parâmetro, a variabilidade não é bem modelada ocasionando altas taxas de falsos positivos (ROBINSON et al., 2010). Assim, visando a redução da taxa de falso positivos, metodologias baseadas na distribuição binomial negativa foram propostas na literatura (BULLARD, 2010).

Dentre as diversas metodologias para o estudo da expressão diferencial, podemos citar o DEGSeq (WANG et al., 2010), baySeq (HARDCASTLE e KELLY, 2009) e DESeq (ANDERS e HUBER, 2010).

2.2.4.1. DEGSeq

O DEGSeq se baseia no MA-plot[5]. Especificamente, o método do DEGSeq pressupõe que o número de reads mapeados (alinhados) a um determinado gene segue distribuição binomial e utiliza a distribuição condicional de M dado A para obter um estatística de teste para avaliar a expressão diferencial entre duas condições experimentais.

2.2.4.1.1. Dedução da distribuição de M|A

Considere que C_1 e C_2 denotam as o número de reads mapeadas para um gene específico obtido a partir de duas amostras, sendo $C_i \sim$ Binomial (n_i, p_i), $i = 1,2$, em que n_i representa o número total de leituras mapeadas e p_i a probabilidade de uma leitura proveniente desse gene. Assumindo que C_1 e C_2 são independentes, o MA-plot é definido por (Wang et. al.,2010):

$$M = X - Y \qquad (01)$$

[5] Gráfico amplamente utilizado para detectar e visualizar razões dependentes da intensidade de luz emitida de dados de microarray. Leitores interessados devem consultar Yang et al. (2002).

e

$$A = (X + Y)/2 \tag{02}$$

em que $X = \log_2(C_1)$ e $Y = \log_2(C_2)$.

De acordo com o Teorema Central do Limite[6] (CASELLA e BERGER, 2002), quando n_1 e n_2 são suficientemente grandes, as distribuições de C_1/n_1 e C_2/n_2 são dadas por:

$$\sqrt{n_1}\left(\frac{C_1}{n_1} - p_1\right) \to N(0, p_1(1-p_1)) \qquad \text{E} \qquad \sqrt{n_2}\left(\frac{C_2}{n_2} - p_2\right) \to N(0, p_2(1-p_2)),$$

.

Uma vez que X é uma função de C_i/n_i, ou seja, $X = g(C_i/n_i) = \log_2(n_1C_1/n_1)$, em que $g(x) = \log_2 n_1 x$, podemos por meio do método Delta[7] (CASELLA e BERGER, 2002), obter a distribuição assintótica de X quando $n_1 \to \infty$:

$$\sqrt{n_1}\left(X - \log_2(n_1 p_1)\right) = \sqrt{n_1}\left(g(C_1/n_1) - g(p_1)\right) \tag{03}$$

$$\to N\left(0, p_1(1-p_1)\left[g'(p_1)\right]^2\right)$$

$$= N\left(0, \left(\frac{1-p_1}{p_1}\right)(\log_2 e)^2\right) \tag{04}$$

Então, voltando a distribuição de X temos:

$$X \to N\left(\log_2(n_1 p_1), \left(\frac{1-p_1}{n_1 p_1}\right)(\log_2 e)^2\right).$$

De maneira similar temos:

$$Y \to N\left(\log_2(n_2 p_2), \left(\frac{1-p_2}{n_2 p_2}\right)(\log_2 e)^2\right).$$

Visando a simplificação de notação, podemos escrever:

$$X \sim N(\mu_X, \sigma_X^2)$$

e

$$Y \sim N(\mu_Y, \sigma_Y^2)$$

Baseando na suposição de independência entre C_1 e C_2 e utilizando o Teorema da Combinação Linear temos:

---

[6] $\sqrt{(n)}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$

[7] $\sqrt{(n)}(g(X_n) - g(\mu)) \xrightarrow{D} N\left(0, \sigma^2\left[g'(\mu)^2\right]\right)$

$$M \sim N\left(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2\right) = N\left(\mu_M, \sigma_M^2\right)$$

e

$$A \sim N\left(\frac{1}{2}(\mu_X + \mu_Y), \frac{1}{4}(\sigma_X^2 + \sigma_Y^2)\right) = N\left(\mu_A, \sigma_A^2\right)$$

Assim, a distribuição condicional de M|A= a, pode ser obtida:

$$M \mid (A = a) \sim N(\mu_M + \rho\frac{\sigma_M}{\sigma_A}(a - \mu_A), \sigma_M^2(1-\rho^2)),$$

em que ρ é o coeficiente de correlação entre M e A.

A covariância entre M e A pode ser obtida por meio de

$$Cov(M, A) = E(MA) - \mu_M\mu_A = \frac{1}{2}E\left(X^2 - Y^2\right) - \frac{1}{2}\left(\mu_X^2 - \mu_Y^2\right) = \frac{1}{2}\left(\sigma_X^2\sigma_Y^2\right). \quad (05)$$

então,

$$\rho = \frac{Cov(M, A)}{\sigma_M\sigma_A} = \frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}. \quad (06)$$

Então

$$E(M \mid A = a) = \mu_X - \mu_Y + 2\frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}(a - \frac{1}{2}(\mu_X + \mu_Y)), \quad (07)$$

e

$$Var(M \mid A = a) = 4\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}. \quad (08)$$

2.2.4.1.2. Teste de expressão diferencial

Considerando um gene qualquer com valores observados de A = a e M = m desejamos verificar as seguintes hipóteses baseando na demonstração acima:

$$\begin{cases} H_0: p_1 = p_2 = p \\ H_1: p_1 \neq p_2 \end{cases}$$

$$\mu_A = \frac{1}{2}(\mu_X + \mu_Y) = \frac{1}{2}\left(\log_2(n_1 p_1)\log_2\left(n_2 p_2\right)\right) \quad (09)$$

Considerando H_0 verdadeira temos:

$$\mu_A = \frac{1}{2}(\mu_X + \mu_Y) = \log_2(n_1 n_2 p^2). \quad (10)$$

Então

$$p = \sqrt{2^{2\mu_A}/(n_1 n_2)}. \quad (11)$$

Substituindo em E(M|A = a) e V (M|A = a), obtemos:

$$\hat{E}(M \mid A = a) = \log_2(n_1)\log_2(n_2), \quad (12)$$

e

$$\hat{Var}(M \mid A = a) = \frac{4(1-\sqrt{2^{2a}/(n_1 n_2)})(\log_2(e))^2}{(n_1+n_2)\sqrt{2^{2a}/(n_1 n_2)}}. \quad (13)$$

De posse dessas quantidades podemos fazer uso do "teste Z" e encontrar o p-valor associado para indicar se o gene é diferencialmente expresso.

$$Z - escore = \frac{\left|m - \hat{E}(M \mid A = a)\right|}{(\hat{V}(M \mid A = a))^{1/2}} \quad (14)$$

Pode-se também traçar um escore limite considerando 4 desvios padrão de M.

2.2.4.2. 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.

2.2.4.2.1. Definição dos modelos

Ao ser considerado casos de análise mais simples, cujo interesse seja comparar duas condições experimentais, por exemplo, as condições A e B. Onde, para cada condição têm-se duas repetições biológicas, ou seja, no total, quatro bibliotecas denotadas por A_1, A_2, B_1, B_2.

Segundo Hardcastle e Kelly (2010) é procedente 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. Contudo, alguns genes podem apresentar expressão diferencial entre as diferentes condições experimentais e assim, os valores dos parâmetros referentes às amostras da condição A (A_1 e A_2) serão diferentes daqueles provenientes das amostras referentes à condição B (B_1 e B_2). Desta forma, podem-se 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 se pressupõe expressão diferencial entre as condições A e B (as amostras são definidas por $\{A_1, A_2\}$ e $\{B_1, B_2\}$, isto é, 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).

2.2.4.2.2. Abordagem bayesiana para inferência sobre expressão diferencial

Suponha um conjunto de dados de contagens com n amostras, $A = \{A_{1, ..., } A_n\}$, tais que os dados observados para um determinado gene é dado pelo seguinte conjunto (HARDCASTLE e KELLY, 2010):

$$D_c = \left\{\left(u_{lc},\ldots,u_{nc}\right),\left(l_1,\ldots,l_n\right)\right\},$$

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

A partir de então, considera-se um modelo M que pode ser definido pelos seguintes conjuntos $\{E_1, ..., E_m\}$. 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:

$$D_c = \left\{\left(u_{ic} : A_i \in E_q\right),\left(l_i : A_i \in E_q\right)\right\}.$$

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 a posteriori do modelo M dado os dados D_c, ou seja,

$$P\left(M \mid D_c\right) = \frac{P\left(D_c \mid M\right)P\left(M\right)}{P\left(D_c\right)} \qquad (15)$$

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.

2.2.4.2.3. Obtenção de $P(D_c \mid M)$

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

$$P(D_c \mid M) = \int P(D_c \mid K, M) P(K \mid M) dK, \tag{16}$$

em que $P(D_c \mid K,M)$ é a distribuição dos dados condicionada ao modelo e ao conjunto de parâmetros $K = \{\theta_1, ..., \theta_m\}$ 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 $K \mid M$. Como discutido em Hardcastle e 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 superdispersão. 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 (AMARAL, 2015).

Visando trabalhar com os dados originais, ou seja, levando em consideração os tamanhos da biblioteca, Hardcastle e 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:

$$P(u_{ic}; l_i, \phi_q, \mu_q) = \frac{\Gamma(u_{ic} + \phi_q^{-1})}{\Gamma(\phi_q^{-1}) u_{ic}!} \left(\frac{1}{1 + l_i \mu_q \phi_q}\right)^{\phi_q^{-1}} \left(\frac{l_i \mu_q}{\phi_q^{-1} + l_i \mu_q}\right)^{u_{ic}}. \tag{17}$$

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 e Kelly (2010) assumiram

que o primeiro parâmetro θ_q, pertencente ao conjunto K, são independentes em relação à q, ou seja,

$$\begin{aligned} P(D_c \mid M) &= \int P(D_c \mid K, M) P(K \mid M) dK \\ &= \prod_q \int P(D_{qc} \mid \theta_q) P(\theta_q) \end{aligned} \tag{18}$$

Essa suposição permite reduzir a dimensionalidade da integral, por consequência, melhora a precisão da sua aproximação numérica. 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 (EVANS e SWARTZ, 1995):

$$P(D_c \mid M) \approx \prod_q \frac{1}{|\Theta_q|} \sum_{\Theta_q} \left[\prod_{\{i: A_i \in E_q\}} \frac{\Gamma(u_{ic} + \Phi_q^{-1})}{\Gamma(\Phi_q^{-1}) u_{ic}!} \left(\frac{1}{1 + l_i \mu_q \Phi_q} \right)^{\Phi_q^{-1}} \left(\frac{l_i \mu_q}{\Phi_q^{-1} + l_i \mu_q} \right)^{u_i} \right]. \tag{19}$$

A partir de então, para obter tal aproximação é necessário obter o conjunto Θ_q com base nos dados adquirindo uma distribuição empírica sobre K.

2.2.4.2.4. 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 e Kelly (2010) o principal problema nessa abordagem se refere à obtenção das estimativas de dispersão, uma vez que não sabemos, previamente, se os genes são ou não diferencialmente expressos, é preciso considerar a estrutura de replicação dos dados, a fim de estimar corretamente as dispersões, isto é, 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 os conjuntos $\{F_1, ..., F_s\}$, onde i, $j \in F_r$ se, e somente se, a amostra A_j for uma réplica de A_i.

Dada esta estrutura para os dados, pode-se estimar a dispersão dos dados num gene D_c pelo método denominado quase probabilidade (NELDER, 2000). Métodos Quasi-verosimilhança são utilizados para dar boas estimativas da dispersão de um único gene nesta configuração (ROBINSON e SMYTH, 2008). Primeiramente defini-se

$$\hat{\mu}_{rc} = \left\langle \left\{ \frac{u_{ic}}{l_i} : i \in F_r \right\} \right\rangle \tag{20}$$

E, em seguida, se escolhe φ_c, tal que

$$2\sum_r \sum_{i \in F_r} \left\{ u_{ic} \log\left[\frac{u_{ic}}{l_i \hat{\mu}_{rc}} \right] - \left(u_{ic} + \Phi_c^{-1} \right) \log\left[\frac{u_{ic} + \Phi_c^{-1}}{l_i \hat{\mu}_{rc} + \Phi_c^{-1}} \right] \right\} = n - 1. \tag{21}$$

Utilizando o valor adquirido para φ_c , pode-se então re-estimar os valores $\hat{\mu}_{ic}$ pelo método da máxima verossimilhança para obter os valores de $\hat{\mu}_{ic}$ que maximizam a seguinte verossimilhança, ou seja, as probabilidades a posteriori para cada q.

$$P\left(\{u_{ic} : i \in F_r\}; l_i \in F_r, \phi_c, \hat{\mu}_{rc} \right) = \prod_{i \in F_r} \frac{\Gamma\left(u_{ic} + \phi_c^{-1} \right)}{\Gamma\left(\phi_c^{-1} \right)} \left(\frac{1}{1 + l_i \hat{\mu}_{rc} \phi_c} \right)^{\phi_c^{-1}} \left(\frac{l_i \hat{\mu}_{rc}}{\phi_c^{-1} + l_i \hat{\mu}_{rc}} \right)^{u_{ic}}, \tag{22}$$

Em seguida, repetir nas estimativas de φ_c e $\hat{\mu}_{ic}$ até alcançar a convergência.

Isso nos dá um valor de φ_c. Em seguida, precisa-se calcular a média da distribuição implícita aos dados D_{qc}, isto é, para o conjunto de amostras no E_q, o que pode facilmente fazer com que o valor adquirido de φ_c seja fixo e calcular a média μ_{qc} pelo método da máxima verossimilhança, escolhendo o valor de μ_{qc} que maximiza a seguinte probabilidade.

$$P\left(D_{qc}, \phi_c, \mu_{qc} \right) = \prod_{\{i : A_i \in E_q\}} \frac{\Gamma\left(u_{ic} + \phi_c^{-1} \right)}{\Gamma\left(\phi_c^{-1} \right) u_{ic}!} \left(\frac{1}{1 + l_i \mu_{qc} \phi_c} \right)^{\phi_c^{-1}} \left(\frac{l_i \mu_{qc}}{\phi_c^{-1} + l_i \mu_{qc}} \right)^{u_{ic}}, \tag{23}$$

para cada q.

A partir disso, forma-se o conjunto $\Theta_q = \{(\mu_{qc}, \varphi_c)\}$, repetindo este processo por múltiplas vezes (h), e são então capazes de calcular

$$P\left(D_c \mid M \right) \approx \prod_q \frac{1}{\left| \Theta_q \right|} \sum_{\Theta_q} \left[\prod_{\{i : A_i \in E_q\}} \frac{\Gamma\left(u_{ic} + \Phi_q^{-1} \right)}{\Gamma\left(\Phi_q^{-1} \right) u_{ic}!} \left(\frac{1}{1 + l_i \mu_q \Phi_q} \right)^{\Phi_q^{-1}} \left(\frac{l_i \mu_q}{\Phi_q^{-1} + l_i \mu_q} \right)^{u_i} \right]. \tag{24}$$

2.2.4.2.5. 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 $p' = \langle P(M|D_c) \rangle_c$. Esse processo deve ser repetido até que o valor de p apresente convergência (AMARAL, 2015).

2.2.4.2.6. 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) (AMARAL, 2015).

2.2.4.2.7. Avaliação das hipóteses

As probabilidades a posteriori, permiti a análise dos testes de hipótese sob o modelo apresentar expressão diferencial com maior valor de probabilidade quando comparado ao modelo que considera que as amostras tenham os mesmos valores paramétricos, diante disso, rejeitar a hipótese da não existência da expressão diferencial. Do contrário, não se deve rejeitar a hipótese de igualdade dos parâmetros.

2.2.4.3. DESeq

O DESeq é um método 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 (ANDERS e HUBER, 2010).

2.2.4.3.1. Descrição do modelo utilizado

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

$$K_{ij} \sim NB(\mu_{ij}, \sigma^2_{ij}), \tag{25}$$

em que K_{ij} é o número de reads de um gene i numa determinada amostra j, μ_{ij} é a média e σ^2_{ij} é a variância.

Na prática os parâmetros μ_{ij} e σ^2_{ij} são desconhecidos, com isso há a necessidade de estimá-los a partir dos dados. Sabendo que o número de repetições em experimentos de RNA-Seq é pequeno, Anders e Huber (2010) propuseram que mais hipóteses de modelagem sejam feitas, a fim de se obter estimativas confiáveis. Este método se desenvolve, então, a partir de três premissas descritas a seguir:

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

$$\mu_{ij} = q_{i,\rho(j)} s_j \tag{26}$$

- A variância σ^2_{ij}, é definida como uma função da média e uma quantidade ajustada da variância amostral suavizada.

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_{i,\rho(j)} \quad (27)$$

- A variação por cada gene com parâmetro $v_{i,\rho}$ é uma função suavizada de q_i e a condição experimental de ρ.

$$v_{i,p(j)} = v_p\left(q_{i,p(j)}\right) \quad (28)$$

A terceira hipótese é necessária porque o número de repetições é geralmente muito pequeno para obter uma estimativa precisa da variação do gene i.

2.2.4.3.2. 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 (AMARAL, 2015):

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, Anders e Huber (2010) utiliza o fator de tamanho s_j. O estimador do fator de tamanho é a mediana de razão das contagens observadas.

$$\hat{s}_j = \underset{i}{\text{mediana}} \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{1/m}} \quad (29)$$

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

Para estimar $q_{i\rho}$, usamos a média das contagens das j amostras correspondentes à condição ρ, transformada para a escala comum por meio do fator de tamanho:

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j}, \tag{30}$$

em que m_ρ é o número de repetições da condição ρ e a soma ocorre sobre estas repetições, e K_{ij} é o nº de reads para o gene i na amostra j.

Observa-se que os valores de k são divididos pelo tamanho do fator correção. Essa estratégia serve para tornar possível a comparação entre diferentes amostras.

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

$$w_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\rho(j)=\rho} \left(\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho} \right), \tag{31}$$

em que

$$z_{i\rho} = \frac{\hat{q}_{i\rho}}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{1}{\hat{s}_j}. \tag{32}$$

A quantidade $z_{i\rho}$ é a correção do viés do estimador da variância bruta. De acordo com Anders e Huber (2010), $w_{i\rho}$ - $z_{i\rho}$ é um estimador não viesado da variância bruta $s_j^2 v_{i,\rho(j)}$. Apesar da utilidade, para os casos em que o número de repetições é pequeno, como nos experimentos de RNA-Seq, a quantidade $w_{i\rho}$ é altamente variável, portanto $w_{i\rho}$ - $z_{i\rho}$ 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 $w_{i\rho}$ e independente $\hat{q}_{i\rho}$ para obter a função suave $w_\rho(q)$ e estimar a variância bruta por meio de (LOADER, 1999):

$$\hat{v}_\rho\left(\hat{q}_{i\rho}\right) = w_\rho\left(\hat{q}_{i\rho}\right) - z_{i\rho}. \tag{33}$$

2.2.4.3.3. 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 (AMARAL, 2015):

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,

$$p = \frac{\binom{N_1}{n_{11}}\binom{N_2}{n_{21}}}{\binom{N..}{n_{11}+n_{21}}} \tag{34}$$

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

$$p\left(reads \geq n_{11}\right) = \sum_{k-n_{11}}^{N_1} \frac{\binom{k+n_{12}}{k}\binom{N_2}{n_{21}}}{\binom{N..}{k+n_{21}}} \tag{35}$$

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, as hipóteses a serem avaliadas, por Andes e Huber (2010), é dada por:

$$H_0 : q_{iA} = q_{iB} \qquad versus \qquad H_1 : q_{iA} \neq q_{iB}$$

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} (AMARAL, 2015),

$$p_i = \frac{\sum_{p(a,b) \leq p(k_{iA}, k_{iB})} p(a,b)}{\sum_{a+b=k_{iS}} p(a,b)} \tag{36}$$

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 e Huber (2010) é dada por:

$$P(K = k) = \binom{k + r - 1}{r - 1} p^{r} (1 - p)^{k}, \tag{37}$$

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

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

2.2.5. Correção de testes múltiplos

Diante da grande quantidade de testes realizados, uma hipótese para cada gene a probabilidade conjunta de que o erro tipo I seja cometido aumenta expressivamente. O erro tipo I, também denominado de falsos positivos (CASELLA, 2010), é o erro que se comete ao rejeitar a hipótese nula quando a mesma é verdadeira (BENJAMINI e HOCHBERG, 1995).

Para controlar a taxa de falsos positivos pode-se utilizar o ajuste de FDR (False Discovery Rate).

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

$$q^{*} \geq \frac{m p_{(i)}}{i} \tag{38}$$

em que q* é o ponto de corte, m é o total das hipóteses testadas, p é o p-valor e i é a ordem do p-valor.

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

A título de ilustração do procedimento, considere-se o exemplo apresentado por Benjamini e Hochberg (1995), em que foram realizados 15 testes de hipóteses, e que os p-valores ordenados, tenham sido: 0,0001; 0,0004; 0,0019; 0,0095; 0,0201; 0,0278; 0,0298; 0,0344; 0,0495; 0,3240; 0,4262; 0,5719; 0,6528; 0,7590 e 1,0000.

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

$$\alpha^* = 0,05 \Rightarrow 0,05 \geq \frac{15P_i}{i} \qquad (39)$$

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

REFERÊNCIAS BIBLIOGRÁFICAS

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

AMARAL, R. T. **Número de repetições na identificação de genes diferencialmente expressos em experimentos de RNA-Seq.** 2015. 60 f. Dissertação (Mestrado em Estatística e Biometria) – Universidade Federal de Viçosa, Viçosa, Minas Gerais, MG, 2015.

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 MolBiol**, 10:Article 26. 2011.

BAILÃO, A. M. **Análises transcricionais no estudo de expressão gênica de Paracoccidioides brasiliensis em condições que mimetizam nichos do hospedeiro**. 2008. 95 f. Tese (Doutorado em Patologia Molecular) – Universidade Federal de Brasília, Brasília, DF, 2008.

BAXTER, E. M., JARVIS, S., D'EATH, R.B., ROSS, D.W., ROBSON, S. K., FARISH, M., NEVISON, I. M., LAWRENCE, A. B., EDWARDS, S. A. Investigating the behavioural and physiological indicators of neonatal survival in pigs. **Theriogenology**. 69: 773–783, 2008.

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

BHERING, L. L.; CRUZ, C. D. Tamanho de população ideal para mapeamento genético em famílias de irmãos completos. **Pesquisa Agropecuária Brasileira**, 43.3, 379-385, 2008.

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. (Versão em português da 2nd edição em inglês), 2010.

CASELLA, G.; BERGER, R. L. **Statistical Inference.** Vol. 2. Pacific Grove, CA: Duxbury, 2002.

CASTAN, E. P. **Transcriptoma do músculo longissimus dorsi de bovinos machos adultos da raça Nelore.** 2014. 80 f. Tese (Doutorado em Genética e Melhoramento) - Universidade

Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, São Paulo, SP, 2014.

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

DIOLA, V. **Resistência à ferrugem do cafeeiro: Mapeamento genético, físico e análise da expressão gênica em resposta a infecção de H. vastatrix.** 2009. 90 f. Tese (Doutorado em Fisiologia Vegetal) – Universidade Federal de Viçosa, Programa de Pós-Graduação em Fisiologia Vegetal, Viçosa, Minas Gerais, MG, 2009.

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

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

FALEIRO, F. G.; ANDRADE, S. R. M.; REIS JUNIOR, F. B. Biotecnologia: o estado da arte e da aplicação a agropecuária. 730 f. **Embrapa Cerrados**, Planaltina, DF, 2011.

FERRAZ, A. L. J. **Ánálise da expressão gênica no músculo esquelético de bovinos das raças nelore e aberdeen angus e sua relação com o desenvolvimento muscular e a maciez da carne.** 2009. 96 f. Tese (Doutorado em Zootecnia) – Universidade Estadual Paulista. Jaboticabal, São Paulo, SP, 2009.

FERREIRA FILHO, D. **Estudo de expressão gênica em citros utilizando modelos lineares.** Dissertação **(**Mestrado em Estatística) - Universidade de São Paulo, São Paulo, SP, 2009.

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

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.

GREENWOOD, P. L.; BELL, A. W.; VERCOE, P. E.; VILJOEN, G. J. Managing the prenatal environment to enhance livestock productivity. **Dordrecht: Springer**, 298p, 2010.

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.

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

LI, J.; TIBSHIRANI, R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data. **Stat Methods Med Res.** apub ahead of print. 2011.

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

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

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

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

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

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

OSHLACK, A.; WAKEFIELD, M. J. Tanscript length bias in RNA-seq data confounds sistems biology. **Genome Biology**. Biology Direct, 4:14, 2009.

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.

SCHENA, M.; SHALON,D.; Davis, R. W.; Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science**, New Series, Vol. 270, No. 5235, pp. 467-470. October, 20, 1995.

SCHUSTER, I.; CRUZ, C.D. **Estatística genômica - aplicada a populações derivadas de cruzamentos controlados.** Viçosa: UFV, 568p. 2004.

SHULTZ, J. L.; KAZI, S; BASHIR, R.; AFZAL, J. A.; LIGHTFOOT, D. A. The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. **Theoretical and Applied Genetics**, 114.6, 1081-1090, 2007.

SILVA, A. P. M. **Expressão gênica associada à resistência da soja a Piezodorusguildinii.** 2014. 119 f. Dissertação (Mestrado em Ciência. Área de concentração: Genética e melhoramento de plantas) – Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, São Paulo, SP, 2014.

SMYTH, G. K. Modelos lineares e métodos de Bayes empíricos para avaliar a expressão diferencial em experimentos de microarranjos. **Dados Appl Genet MolBiol**, a 3: Artigo 3, 2004.

SOLERO, B. P. Transcriptional profiling during foetal skeletal muscle development of Piau and Yorkshire – Landrace cross-bred pigs. **Animal Genetics**, v.42, n. 6, p. 600-612, nov. 2010.

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

SULTAN, M.; SCHULZ, M. H.; RICHARD, H.; MAGEN, A.; KLINGENHOFF, A.; SCHERF, M.; SEIFERT, M.; BORODINA, T,; SOLDATOV, A.; PARKHOMCHUK, D.; SCHMIDT, D.; O'KEEFFE, S.; HAAS, S.; VINGRON, M.; LEHRACH, H.; YASPO, M. L. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. **Ciência**, 321:. 956-960, 2008.

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

YANG, Y. H.; DUDOIT, S. D.; LUU, P.; SPEED, T. P. Normalization for cDNA Microarray Data. **In SPIE BioE**, 2001.

YANG, Y. H.; DUDOIT, S.; LUU, P.; LIN, D. M.; PENQ, V.; NQAI, J.; SPEED, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. **Nucleic Acids Res.**, 30, e15, 2002.

CAPÍTULO 2

Metodologias para identificação de genes diferencialmente expressos em experimentos de RNA-Seq de suínos

Resumo

O objetivo deste estudo é avaliar o comportamento de três metodologias (DEGSeq, bayseq e DESeq) para verificação da expressão diferencial em longissimus dorsi (LD) do músculo de suínos da raça Piau e Comercial, em 21e 90 dias depois do coito, por meio de dados provenientes de RNA-Seq. Para tanto, foram utilizados dados de 30000 genes provenientes do músculo longissimus dorsi (LD), coletados de dois grupos genéticos de suínos (Piau e uma linha comercial desenvolvida na UFV), sob duas condições experimentais (idades pós-natal), 21 e 90 dias depois do coito. De acordo com os resultados gerados nas análises e sob as condições utilizadas no desenvolver do experimento, pôde-se concluir, na comparação dos métodos bayseq com DEGSeq e baySeq com DESeq, respectivamente, observou-se, a partir da relação do nível de expressão (fold-change) entre as duas raças suínas (comercial e piau), que os métodos apresentaram desempenho diferentes entre si, pois apresentaram um nível de expressão desigual em ambos os métodos. No entanto, na comparação entre os métodos DESeq e DEGSeq, houve um desempenho comparável, isto é, os métodos são considerados concordantes em relação aos resultados. Como um todo, a maioria dos genes DE identificados, se deu na fase pós-natal tardia, ou seja, 90 dpc. Além disso, a maioria deles foram down na fase pré-natal inicial (21 dpc) e foram up na fase pré-natal tardia (90 dpc) relacionando as raças, comercial e piau e comparando os métodos.

Palavras-chave: DEGSeq, baySeq, DESeq, bioinformática.

1. INTRODUÇÃO

O estudo do desenvolvimento muscular na fase pré-natal de suínos é fundamental importância para o entendimento de características de importância econômica, tais como a sobrevivência e peso ao nascer (BAXTER et al., 2008). Segundo GREENWOOD et al. (2010) tal etapa é considerada um período de desenvolvimento máximo dos tecidos e órgãos, sendo influenciado tanto por fatores genéticos quanto ambientais. Solero et al. (2011), por meio de experimentos de microarray, identificaram genes que apresentam expressão diferencial em longissimus dorsi (LD) do músculo de suínos em 40 e 70 dias de gestação em suínos da raça Piau e Yorkshire-Landrace.

Apesar de útil, a técnica de microarray utilizada nesse estudo, apresenta algumas limitações quando comparada ao novo método sequenciamento de RNA em larga escala (RNA-Seq). Dentre essas limitações, pode-se citar a necessidade do conhecimento prévio das sequências, a expressão de um gene é mensurada de maneira relativa por meio da intensidade de luz emitida em cada condição de interesse, problema de susceptibilidade à hibridização cruzada, baixa reprodutibilidade de resultados entre laboratórios e diferentes plataformas e ineficiência no estudo de genes raros e de isoformas (ESTEVES, 2007; FERREIRA FILHO, 2009; NEVES, 2010). Assim, o estudo do desenvolvimento muscular na fase pré-natal em suínos por meio de dados provenientes da técnica de RNA-Seq torna-se interessante.

A literatura apresenta diversas metodologias para o estudo da expressão diferencial (ED), as quais diferem quanto aos conceitos adotados e aos princípios estatísticos empregados. Como exemplo, pode-se citar as metodologias DEGSeq (WANG et al., 2010), baySeq (HARDCASTLE e KELLY, 2009) e DESeq (ANDERS e HUBER, 2010).

Diante da grande quantidade de metodologias e da falta de um consenso sobre qual a melhor para avaliar ED, o objetivo deste estudo é avaliar o comportamento de três metodologias (DEGSeq, baySeq e DESeq) para verificação da expressão diferencial em músculo longissimus dorsi (LD) de suínos da raça Piau e Comercial, em 21e 90 dias depois do coito (dpc), por meio de dados provenientes de RNA-Seq.

2. Material e Métodos

Com o intuito de avaliar os métodos de identificação de expressão diferencial, foram analisados 30000 genes do músculo longissimus dorsi (LD), coletados de dois grupos genéticos de suínos (Piau e uma linha comercial desenvolvida na UFV), sob duas condições experimentais (idades pós-natal), 21 e 90 dias depois do coito, que correspondem a situações sem repetição. Utilizou-se o kit RNease Mini (QIAGEN) para a extração das amostras, cinco

microgramas (5 µg) de RNA total foi utilizado para iniciar a construção da biblioteca e análise transcricional. Para a obtenção dos fetos, foram abatidas fêmeas nas respectivas idades gestacionais de acordo com as normas do Comitê de Ética na Experimentação animal da UFV.

Visando comparar as metodologias, foi calculada a quantidade de genes diferencialmente expressos identificados em cada método, para verificar o poder de quantificação entre eles. Posteriormente, com o intuito estudar a sobreposição entre os conjuntos de genes DE, ou seja, genes que foram compartilhados com outros métodos construiu-se o diagrama de venn.

Finalmente, para avaliar o comportamento dos genes diferencialmente expressos, foram ajustados modelos de regressão linear simples para cada idade (21 e 90 dias depois do coito):

$$Y = \beta_0 + \beta_1 X + e, \tag{1}$$

em que

- Y = fold-change (baySeq) e X = fold-change (DESeq)
- Y = fold-change (DESeq) e X = fold-change (DEGSeq)
- Y = fold-change (DEGSeq) e X = fold-change (baySeq)

Onde, o valor do fold-change, ou seja, o valor da razão de expressão em logaritmo de base 2 (log2) da raça comercial em relação a raça piau, em cada idade, é calculado:

$$\text{Fold} - \text{Change} = \log_2\left(\frac{C_{ij}}{P_{ij}}\right), \tag{2}$$

em que C_{ij} é a expressão, na raça comercial, do gene i na j-ésima amostra e P_{ij} é a expressão, na raça piau, do gene i na j-ésima amostra.

3. Resultados e Discussão

O método baySeq apresentou um maior número de genes diferencialmente expressos quando comparados com os demais métodos utilizados nesse estudo, Esse resultado é corroborado com aqueles obtidos por Soneson e Delorenzi (2013) em que os autores mostraram que o método DESeq tem uma abordagem conservadora, já que em geral superestima a variância para cada gene, diminuindo a identificação de genes diferencialmente expressos. Além disso, pode observar que o DESeq apresenta um poder inferior quando comparado ao bayseq e DEGseq.

Tabela 1: Quantidade de genes diferencialmente expressos identificados simultaneamente, em cada método, por idade (em dias depois do coito).

Idade	Genes Diferencialmente Expressos		
	DESeq	baySeq	DEGSeq
21	165	12879	8712
90	265	10217	7666

Ao estudar a sobreposição entre os conjuntos de genes DE entre os três métodos, pôde-se verificar que 94,5455% dos genes DE pelo método DESeq são também DE pelo método DEGSeq, que 55,2916% dos genes DE pelo método DEGSeq são também DE no método baySeq e por fim, que 0,8463% dos genes DE pelo método bayseq foram também DE no método DESeq (Figura 1A). Além disso, verifica-se que 89,4340% dos genes DE pelo método DESeq são também DE pelo método DEGSeq, que 61,2314% dos genes DE pelo método DEGSeq são também DE no método baySeq e por fim, que 2,5056% dos genes DE pelo método bayseq foram também DE no método DESeq (Figura 1B). Em ambas as figuras, baySeq e DEGSeq encontraram uma boa quantidade de genes DE que não foram compartilhados com outros métodos, certificando que o DESeq apresenta um poder inferior quando comparado ao bayseq e DEGseq.

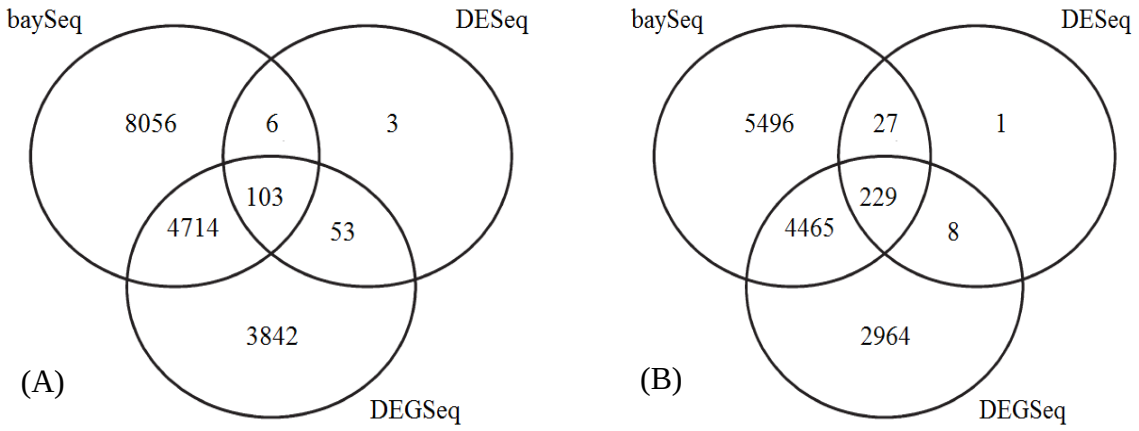

Figura 1. Diagrama de Venn dos Números dos Genes Diferencialmente Expressos entre os Métodos, em 21 dias (A) e 90 dias (B) depois do Coito, com FDR menor que 5%.

Observa-se maior número de genes DE em 90 dpc, ao se comparar DESeq e baySeq, e também, um coeficiente de correlação forte sob os valores do fold-change (relacionando os níveis de expressão gênica das raças comercial e piau), indicando que há relação positiva entre os níveis de expressão. Entre DESeq e DEGSeq, observou-se um maior número de genes DE em 90 dpc, e em ambas as idades, 21 e 90 dpc, os valores do fold-change são altamente correlacionados. Além disso, entre os métodos DEGSeq e baySeq, o maior número de genes DE obteve-se em 21dpc, porém os genes com idade de 90 dpc, apresentaram um

coeficiente de correlação forte entre os valores do fold-change, mostrando que há relação entre os níveis de expressão. Com tudo, a maioria dos genes DE identificados, se deu na fase pré-natal tardia, ou seja, 90 dpc, reforçando os resultados de Cagnazzo et al. (2006), que examinaram a expressão do gene no desenvolvimento de músculo esquelético de duas raças de porcos que diferem em características das fibras musculares e fenótipos de muscularidade. Eles descobriram que vários genes relacionados com a miogênese foram expressos em sua fase pré – natal tardia.

Tabela 2: Quantidade de genes diferencialmente expressos identificados simultaneamente, entre os métodos, com os coeficientes de correlação e os modelos de regressão, por idade (em dias depois do coito).

Comparação	Idade (Comercial versus Piau)	Números de Genes DE	Coeficiente de Correlação	Modelos de Regressão
baySeq e DESeq	21	109	0,6511	Y = 0,1837 + 0,4462x
	90	256	0,9729	Y = 0,3216 + 0,8939x
DESeq e DEGSeq	21	156	1,0000	Y = 0,1824 + 1,0000x
	90	237	1,0000	Y= - 0,1885 +1,0000x
DEGSeq e baySeq	21	4817	0,2706	Y = - 0,3598 + 0,1640x
	90	4694	0,9095	Y = - 0,0717 + 0,8671x

Constatou-se na figura 2, que na fase inicial, ou seja, na idade de 21 dpc, dos 109 genes DE identificados entre os dois métodos, 30% desses genes obtiveram o valor positivo do fold-change, ou seja, o nível de expressão foi maior na raça comercial, esses então, são up em baySeq e DESeq; 56% apresentaram valor negativo do fold-change, ou seja, o nível de expressão foi maior na raça piau, sendo down em baySeq e DESeq. Porém ocorreu uma expressão de maneira diferente já que os valores do fold-change não se ajustaram completamente à equação de regressão representada pela expressão gênica nos métodos baySeq (X) e DESeq (Y). Isto é, pode-se considerar que os métodos são diferentes, pois o nível de expressão foi desigual em ambos. Além disso, 14% dos genes DE apresentaram resultados discordantes, isto é, obtiveram níveis de expressão alto e baixo na comparação de ambos os métodos.

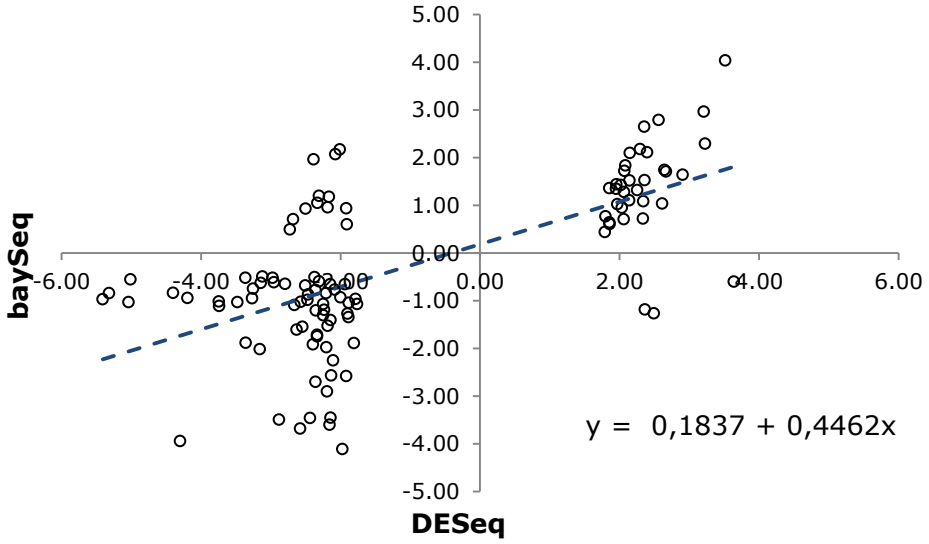

Figura 2. Comportamento dos Genes Diferencialmente Expressos, entre os Métodos DESeq e baySeq, em 21 dias depois do Coito.

Considerando a análise para 90 dpc observa-se que dos 256 genes DE identificados pelos dois métodos, 75% apresentaram o valor positivo do fold-change, ou seja, o nível de expressão foi maior na raça comercial, sendo assim, são up em baySeq e DESeq. Além disso, 25%, dos 256 genes DE, obtiveram o valor negativo do fold-change, ou seja, o nível de expressão foi maior na raça piau, sendo down em baySeq e DESeq. No entanto, o nível de expressão ocorreu de maneira diferente, já que os valores do fold-change não se ajustaram completamente à equação de regressão representada pela expressão gênica nos métodos baySeq (X) e DESeq (Y), logo, podendo considerar que os métodos são diferentes já que a expressão foi desigual em ambos (Figura 3).

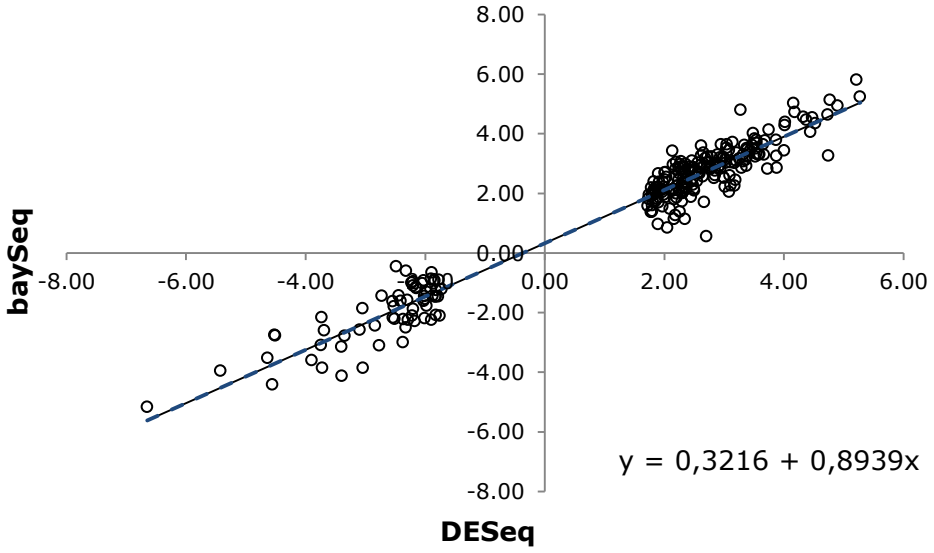

Figura 3. Comportamento dos Genes Diferencialmente Expressos, entre os Métodos DESeq e baySeq, em 90 dias depois do Coito.

Pôde-se observar na figura 4, que dos 156 genes DE identificados por ambos os métodos, 21% apresentaram o valor positivo do fold-change, ou seja, o nível de expressão foi maior na raça comercial, sendo esses, up em DEGSeq e DESeq. Além do mais, 79%, dos 156 genes DE, obtiveram o valor negativo do fold-change, ou seja, o nível de expressão foi maior na raça piau, sendo down em DEGSeq e DESeq. Constatando que, a expressão up e down ocorreu de maneira semelhante, já que os valores do fold-change se ajustaram completamente à equação de regressão representada pela expressão gênica nos métodos DESeq (X) e DEGSeq (Y), logo, ao apresentarem níveis de expressão semelhantes, os métodos são considerados concordantes em relação aos resultados.

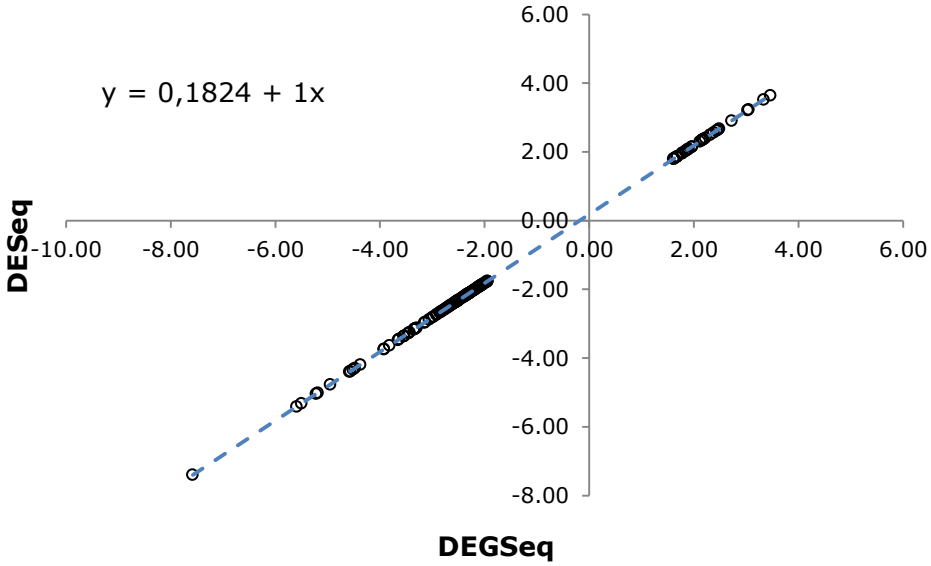

Figura 4. Comportamento dos Genes Diferencialmente Expressos, entre os Métodos DEGSeq e DESeq, em 21 dias depois do Coito.

Na Figura 5, observou-se que dos 237 genes DE identificados por ambos os métodos, 75% apresentaram o valor positivo do fold-change, ou seja, o nível de expressão foi maior na raça comercial, logo esses genes são up em DEGSeq e DESeq. E 25%, dos 237 genes DE, obtiveram o valor negativo do fold-change, ou seja, o nível de expressão foi maior na piau, sendo down em DEGSeq e DESeq. Constatando que, a expressão up e down ocorreu de maneira semelhante, já que os valores do fold-change se ajustaram completamente à equação de regressão representada pela expressão gênica nos métodos DESeq (X) e DEGSeq (Y), logo, ao apresentarem níveis de expressão semelhantes, os métodos são considerados concordantes em relação aos resultados.

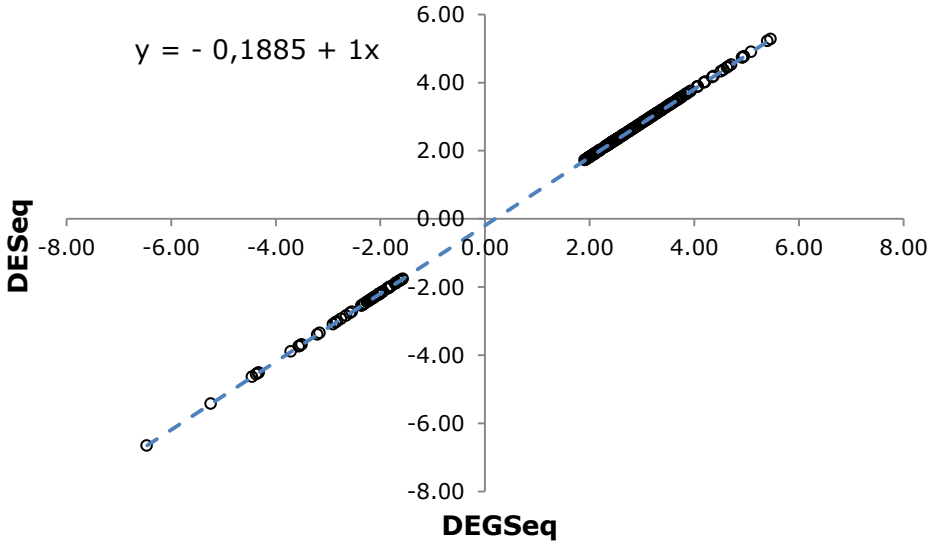

Figura 5. Comportamento dos Genes Diferencialmente Expressos, entre os Métodos DEGSeq e DESeq, em 90 dias depois do Coito.

Comparando os métodos baySeq e DEGSeq, em 21 dpc. Observou-se que, dos 4817 genes DE identificados pelos métodos baySeq e DEGSeq, 14% obtiveram o valor positivo do fold-change, ou seja, o nível de expressão foi maior na raça comercial e são up em baySeq e DEGSeq; 54% dos genes DE apresentaram o valor negativo do fold-change, ou seja, o nível de expressão foi maior na raça piau e são down em baySeq e DEGSeq, porém uma expressão de forma diferente, já que os valores do fold-change não se ajustaram completamente à equação de regressão representada pela expressão gênica nos métodos DEGSeq (X) e baySeq (Y). Isto é, pode-se considerar que os métodos são diferentes, pois o nível de expressão foi desigual em ambos. Além disso, 32% dos genes DE apresentaram resultados discordantes, isto é, obtiveram níveis de expressão alto e baixo na comparação de ambos os métodos (Figura 6).

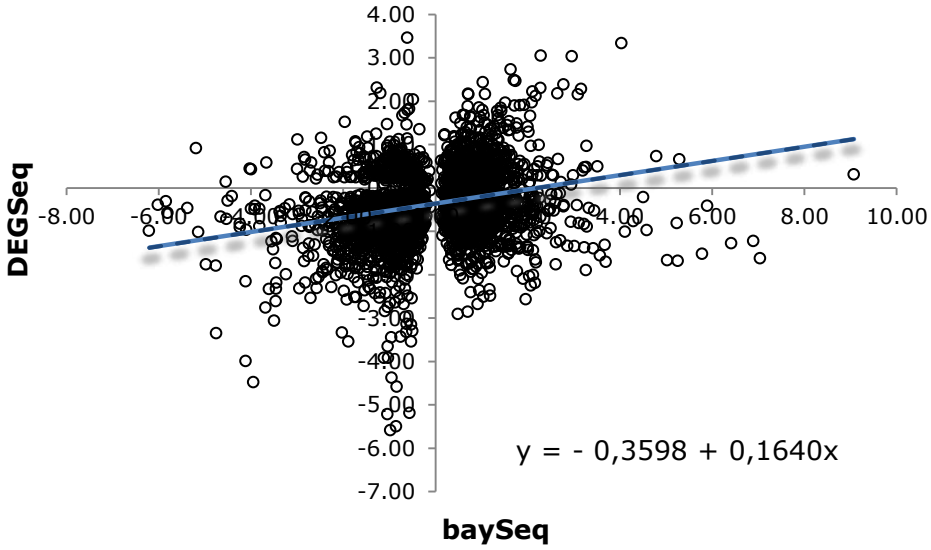

Figura 6. Comportamento dos Genes Diferencialmente Expressos, entre os Métodos baySeq e DEGSeq, por Idade, em 21 dias depois do Coito.

Constatou-se, na figura 7, que dos 4694 genes DE identificados pelos métodos baySeq e DEGSeq, 83% apresentaram o valor positivo do fold-change, ou seja, o nível de expressão foi maior na raça comercial, sendo up nos dois métodos; 14% dos genes DE obtiveram o valor negativo do fold-change, ou seja, o nível de expressão foi maior na raça piau e são down em baySeq e DEGSeq, porém ocorreu uma expressão de maneira diferente, já que os valores do fold-change não se ajustaram completamente à equação de regressão representada pela expressão gênica nos métodos DEGSeq (X) e baySeq (Y). Com isso, pode-se considerar que os métodos são diferentes, pois o nível de expressão foi desigual em ambos. Além disso, 3% dos genes DE apresentaram resultados discordantes, ou seja, obtiveram níveis de expressão alto e baixo na comparação dos métodos.

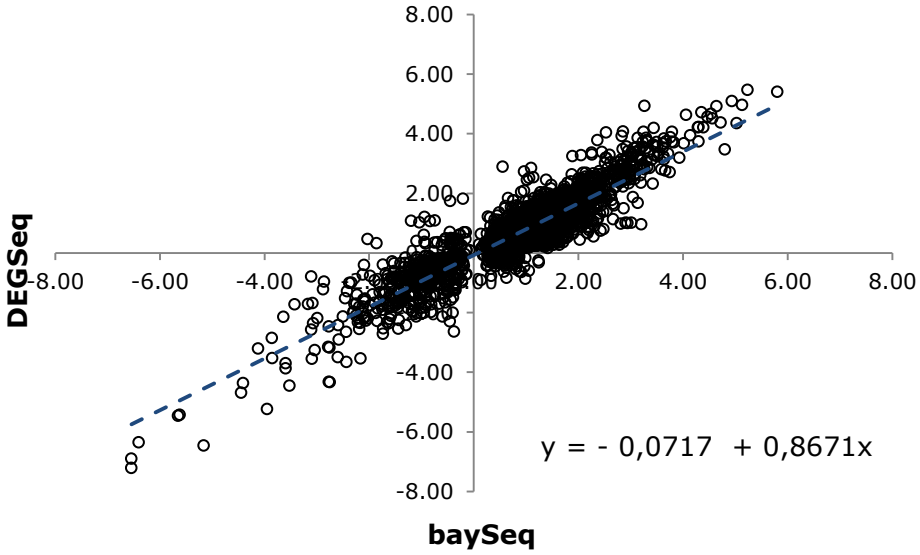

Figura 7. Comportamento dos Genes Diferencialmente Expressos, entre os Métodos baySeq e DEGSeq, por Idade, em 90 dias depois do Coito.

4. Conclusão

- **Comportamento dos métodos**

O método abordado pelo baySeq apresentou um bom desempenho, pois obteve uma maior identificação dos genes diferencialmente expressos, e que o método DESeq apresentou um poder inferior na identificação de genes DE quando comparado ao bayseq e DEGseq.

Ao comparar bayseq com DEGSeq e baySeq com DESeq, respectivamente, observou-se, a partir da relação do nível de expressão, os métodos apresentaram desempenho diferentes entre si. No entanto, na comparação entre os métodos DESeq e DEGSeq, os métodos apresentaram desempenho semelhante, ou seja, os métodos são considerados concordantes em relação aos resultados.

- **Comportamento dos métodos em cada idade**

A maioria dos genes DE identificados se deu na fase pré-natal tardia, ou seja, 90 dpc. Além disso, a maioria dos genes DE foram down na fase pré-natal inicial (21 dpc), com 56% entre baySeq e DESeq, 79% entre DESeq e DEGSeq, e 54% entre DEGSeq e baySeq. Sendo também, a maioria dos genes DE, up na fase pré-natal tardia (90 dpc), com 75% entre baySeq e DESeq, 75% entre DESeq e DEGSeq, e 83% entre DEGSeq e baySeq relacionando as raças, comercial e piau.

5. Considerações Finais

Sob as duas condições utilizadas (21 e 90 dias depois do coito) no desenvolver do experimento, houve uma grande variabilidade nos resultados obtidos pelos três métodos. O método abordado pelo baySeq apresentou um bom desempenho, pois obteve uma maior identificação dos genes diferencialmente expressos, e que o método DESeq apresentou um poder inferior na identificação de genes DE quando comparado ao bayseq e DEGseq.

Comparando bayseq com DEGSeq e DESeq, respectivamente, observou-se, a partir da relação do nível de expressão (fold-change) entre as duas raças suínas (comercial e piau), que os métodos apresentam desempenho diferentes entre si, pois apresentaram um nível de expressão desigual em ambos os métodos. No entanto, na comparação entre os métodos DESeq e DEGSeq, houve um desempenho comparável, isto é, o nível de expressão foi considerado semelhante, logo os métodos são concordantes em relação aos resultados.

A maioria dos genes DE identificados se deu na fase pré-natal tardia, ou seja, 90 dpc. Além disso, a maioria dos genes DE foram down na fase pré-natal inicial (21 dpc), com 56% entre baySeq e DESeq, 79% entre DESeq e DEGSeq, e 54% entre DEGSeq e baySeq. Sendo também, a maioria dos genes DE, up na fase pré-natal tardia (90 dpc), com 75% entre baySeq e DESeq, 75% entre DESeq e DEGSeq, e 83% entre DEGSeq e baySeq relacionando as raças, comercial e piau.

Por fim, a comparação entre os métodos DEGSeq, baySeq e DESeq, a partir de dados provenientes de RNA-Seq, foi capaz de quantificar a expressão gênica diferencial e o comportamento dos métodos, analisando transcriptomas do músculo longissimus dorsi (LD) de suínos da raça Piau e Comercial,em 21 e 90 dias depois do coito.

REFERÊNCIAS BIBLIOGRÁFICAS

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

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.

CAGNAZZO, M.; PAS, M. F. W.; PRIEM, J.; WIT, A. A. C.; POOL, M. H.; DAVOLI, R.; RUSSO, V. Comparison of prenatal muscle tissue expression profiles of two pig breeds differing in muscle characteristics. **J Anim Sci**. 2006 Jan;84(1):1-10, 2006.

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.

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

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.

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.

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

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

WANG, L.; FENG, Z.; WANG, X.; WANG, X.; ZHANG, X. 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.