XII Simpósio Brasileiro de Melhoramento Animal
Ribeirão Preto, SP – 12 e 13 de junho de 2017

Avaliação de abordagens estatísticas para análise da expressão gênica diferencial em dados reais de RNA-seq

Ana Paula Sbardella[1]*, Isabela Fonseca[2], Mayara Morena Del Cambre Amaral Weller[2], Nedenia Bonvino Stafuzza[1], Jaqueline Rosa Oliveira[1], Rafael Nakamura Watanabe[1], Rebeka Magalhães da Costa[1], Alejandro Barrera Carvajal [1], Marcos Vinícius Gualberto Barbosa da Silva[2], Marta Fonseca Martins[2], Danísio Prado Munari[1]

[1]Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, SP, Brasil.
[2] Embrapa Gado de Leite/Juiz de Fora-MG, Brasil.
*Autor correspondente: paulasbardella@gmail.com.br

Resumo: O objetivo deste trabalho foi avaliar três abordagens estatísticas distintas quanto a capacidade de detecção da expressão gênica diferencial de dados obtidos por sequenciamento de RNA extraído de glândula mamária mantida em um sistema extracorpóreo de bovinos de leite mestiços Holandês-Zebu e infectada experimentalmente com *Streptococcus agalactiae*. Úberes de quatro vacas abatidas foram colhidos, perfundidos e inoculados com *S. agalactiae*. Para cada úbere, dois quartos foram inoculados e dois quartos foram utilizados como controle. Biópsias do tecido alveolar foram feitas nos tempos 0 e 3 horas após a perfusão do tecido. O RNA das amostras foi extraído e sequenciado com a plataforma HiSeq2000 (Illumina). A partir das leituras obtidas, os transcritos foram montados utilizando como referência a anotação do genoma bovino UMD 3.1 e as abordagens estatísticas edgeR, baySeq e Cuffdiff foram utilizadas para análise de expressão gênica diferencial. Foram identificados 1756, 1161 e 3389 genes com expressão gênica diferencial pelo edgeR, baySeq e Cuffdiff, respectivamente. As abordagens estatísticas estudadas apresentam grandes diferenças na identificação de genes, resultante dos diferentes parâmetros em que são baseadas. Maior abrangência na identificação de genes associados à infecção da glândula mamária foi obtida com a abordagem Cuffdiff.

Palavras–chave: expressão diferencial, bovinos de leite, sequenciamento.

Comparison of statistical methods for analysis of differential gene expression in real RNA-sequence data

Abstract: The aims of this study was evaluate three different statistic approaches regarding to differential gene expression data obtained by RNA sequencing from mammary glands of Holstein-Zebu crossbred dairy cattle, kept in an extracorporeal system and experimentally infected with *Streptococcus agaclatiae*. Udders of four slaughtered cows were collected, perfused and inoculated with *S. agalactiae*. For each udder, 2/4 were inoculated and 2/4 were used as control. Biopsies were obtained from the alveolar tissue at 0 and 3 hours after the tissue perfusion. The RNA sample was extracted and sequenced with HiSeq 2000 platform (Illumina). From the reads obtained transcripts were assembled using as the reference the bovine genome annotation UMD 3.1 and the statistics approaches edgeR, baySeq and Cuffdiff were used for differential gene expression analysis. 1756, 1161 and 3389 genes with differential gene expression were identified by edgeR, baySeq and Cuffdiff, respectively. The statistical approaches studied present great differences in the identification of genes, resulting from the different parameters on which they are based. Larger comprehensiveness in the identification of genes associated with infection of the mammary gland was obtained with the Cuffdiff approach.

Keywords: differential expression, milk cattle, sequencing.

Introdução

Significativas perdas econômicas em animais de produção são devido a diversas doenças, como infecção da glândula mamária em bovinos leiteiros. Análise de transcriptoma com a utilização da técnica de sequenciamento de RNA, seguida de análise estatística possibilita identificar a expressão gênica diferencial em diversos tecidos infectados experimentalmente. Os resultados permitem a geração de estratégias de manejo e de seleção que priorizem a utilização de animais mais resistentes. O objetivo deste trabalho foi avaliar três abordagens estatísticas distintas quanto a capacidade de detecção da expressão gênica diferencial de dados obtidos por sequenciamento de RNA extraído de glândula mamária mantida em um sistema extracorpóreo de bovinos de leite mestiços Holandês-Zebu e infectada experimentalmente com *Streptococcus agalactiae*.

Material e Métodos

Úberes de quatro vacas abatidas foram colhidos, perfundidos e inoculados com *S. agalactiae*. Para cada úbere, dois quartos foram inoculados (anterior esquerdo - AE; e posterior esquerdo - PE) e dois quartos foram utilizados como controle (anterior direito - AD; e posterior direito - PD). Biópsias foram feitas do tecido alveolar nos tempos 0 e 3 horas após a perfusão do tecido. O RNA das amostras foi extraído e sequenciado com a plataforma HiSeq2000 (Illumina). As sequências curtas de cada amostra foram alinhadas com a versão do genoma bovino referência UMD 3.1, utlizando o algoritmo TopHat. A partir dos resultados do alinhamento das "reads" com o genoma referência bovino, a montagem individual dos transcritos para cada amostra foi realizada utilizando o programa Cufflinks e então as abordagens estatísticas edgeR (ROBINSON et al., 2010), baySeq (HARDCASTLE; KELLY, 2010) e Cuffdiff (TRAPNELL et al., 2013) foram utilizadas para análise de expressão gênica diferencial.

Resultados e Discussão

Foram detectados 1756, 1161 e 3389 genes diferencialmente expressos pelo edgeR, baySeq e Cuffdiff, respectivamente. Apenas 122 genes identificados foram comuns às três abordagens. As abordagens edgeR e baySeq são mais próximas na detecção de genes uma vez que 900 genes encontrados coincidem pelas duas abordagens. Quando utilizadas as abordagens edgeR e baySeq em relação ao Cuffdiff, apenas 221 e 145 genes coincidem, respectivamente.

A diferença no número de genes identificados por cada método é explicada pelos diferentes parâmetros nos quais cada método é embasado (Tabela 1). O método implementado pelo edgeR, utiliza teste exato baseado em distribuição binomial negativa e foi desenvolvido para analisar experimentos com poucas repetições. Este é utilizado para ajustar a sobredispersão ao redor de genes pelo uso da informação de outros genes (ROBINSON; SMYTH, 2008), sendo que a estratégia padrão implementada converge estimativas de dispersão de genes em direção a uma estimativa comum (ROBINSON; McCARTHY; SMYTH, 2010) encontrando o máximo da função.

O método implementado no baySeq assume que a distribuição dos dados é binomial negativa e estima probabilidades posteriores da expressão diferencial por métodos bayesianos empíricos em vez de níveis de significância (HARDCASTLE; KELLY, 2010), encontrando-se as distribuições dos parâmetros. Esse método é teoricamente mais restrito comparado com os demais, o que pode ser comprovado pelos resultados obtidos, dado que foi verificado um número menor de genes diferencialmente expressos para esta metodologia. O método bayesiano pondera os resultados esperados em termos de número de genes significativamente expressos pelas informações a priori.

O método implementado pelo Cuffdiff mede a expressão de um transcrito assumindo distribuição beta binomial negativa para os dados de contagem, sendo que a mudança nessa contagem é considerada para identificar genes diferencialmente expressos (TRAPNELL et al., 2013). Para controlar a variabilidade em profundidade do sequenciamento (OSHLACK; ROBINSON; YOUNG, 2010), o modelo de Poisson é o mais simples e estima a variabilidade por cálculo de contagem média entre réplicas (TRAPNELL et al., 2013).

Os métodos implementados por edgeR e Cuffdiff apresentam suporte para análise pareada dos fatores experimentais e para detecção de genes diferencialmente expressos sem amostras replicadas, todavia, o baySeq não apresenta suporte para esses experimentos (RAPAPORT et al., 2013). Segundo Hardcastle e Kelly (2010), o baySeq tem performance igual ao método edgeR para pequeno número de

bibliotecas, conforme o número bibliotecas aumenta, baySeq demonstra melhora na performance sobre o edgeR.

Tabela 1. Métodos estatísticos para diferencial de expressão gênica

Método	Normalização	Distribuição	Teste de Diferencial de Expressão	Referência
edgeR	TMM[1]	Binomial Negativa	Teste exato	Robinson et al., 2010
baySeq	RLE[2]	Binomial Negativa	Avalia probabilidades posteriores por métodos bayesianos empiricos e os compara	Hardcastle; Kelly, 2010
Cuffdiff	Geométrico e classificado pelo FPKM[3]	Beta Binomial Negativa	Teste *t*	Trapnell et al., 2013

[1]TMM: *trimmed mean of M-values*; [2]RLE: *relative log expression*; [3]FPKM: fragmentos por kilobase de transcritos por milhão de fragmentos mapeados.

Conclusão

As abordagens estatísticas estudadas apresentam grandes diferenças na identificação de genes, que é dependente do experimento biológico e resultante dos diferentes parâmetros em que são baseadas. Para este delineamento experimental as três abordagens estudadas apresentaram bons resultados, sendo que maior abrangência na identificação de genes associados à infecção da glândula mamária foi obtido com a abordagem Cuffdiff.

Agradecimentos

À CAPES-EMBRAPA, FAPEMIG e CNPq.

Literatura citada

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

HARDCASTLE, T.J.; KELLY, K.A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. **BMC Bioinformatics**, v. 11, n. 422, 2010.

TRAPNELL, C.; HENDRICKSON, D.G.; SAUVAGEAU, M.; GOFF, L.; RINN, J.L.; PACHTER, L. Differential analysis of gene regulation at transcript resolution with RNA-seq. **Nature Biotechnology**, v. 31, n. 1, p. 46-53, 2013.

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

OSHLACK, A.; ROBINSON, M.D.; YOUNG, M.D. From RNA-seq reads to differential expression results. **Genome biology**, v. 11, n. 12, p. 220, 2010.

RAPAPORT, F.; KHANIN, R.; LIANG, Y.; PIRUN, M.; KREK1, A.; ZUMBO, P.; MASON, C.E.; SOCCI, N.D.; BETEL, D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. **Genome Biology**, v. 14, n. 9, p. R95, 2013.