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

Regiane Teodoro do AMARAL[1]
Moysés NASCIMENTO[1]
Talles Eduardo Ferreira MACIEL[2]
Fabyano Fonseca e SILVA[2]
Ana Carolina Campana NASCIMENTO[1]
Luiz Alexandre PETERNELLI[1]
Simone Eliza Facioni GUIMARÃES[2]

- RESUMO: Este trabalho teve por objetivo verificar a influência do número de repetições para duas importantes metodologias estatísticas, DESeq e baySeq, usadas na identificação de genes diferencialmente expressos (GDE). Para tanto, considerou-se quatro cenários mimetizando experimentos reais, representando duas condições experimentais, com diferentes números de repetições. O pacote TCC implementado no Bioconductor foi utilizado para simulação de 1000 genes, 200 são diferencialmente expressos (DE) e 800 não DE's. Em um primeiro momento, os dados foram analisados separadamente por cada método, comparando-se a influência do número de repetições na identificação dos GDE. Em seguida, foi realizada a comparação entre os resultados obtidos por cada método, levando em conta também o número de repetições em cada cenário. Observou-se, para os dois métodos analisados, que a redução do número de repetições diminui o poder destas metodologias em detectar GDE. E, que um maior percentual de GDE comuns foi encontrado ao considerar maior número de repetições. O método implementado no pacote baySeq apresentou melhor acurácia nos cenários onde havia 5 repetições e no cenário sem repetições. Assim, pode-se dizer que o baySeq apresentou maior sensibilidade, ou seja, maior taxa de verdadeiros positivos e menor taxa de falsos positivos em relação ao DESeq sob estas condições.

- PALAVRAS-CHAVE: Transcriptoma; simulação; baySeq; DESeq.

1 Introdução

O sequenciamento de RNA (RNA-Seq) tornou-se em pouco tempo a ferramenta mais indicada para estudos de transcriptoma em larga escala (MOROZOVA *et al.*, 2009, WANG *et al.*, 2009), devido principalmente, a maior sensibilidade em relação às técnicas anteriores destinadas a este propósito, tais como *microarrays* e análise seriada da expressão gênica (SEYEDNASROLLAH *et al.*, 2013). Por anos, a análise de *microarrays*

[1] Universidade Federal de Viçosa - UFV, Departamento de Estatística, CEP: 36570-900, Viçosa, MG, Brasil. E-mail: *teodoroamaral@gmail.com; moysesnascim@ufv.br; ana.campana@ufv.br; peternelli@ufv.br*
[2] Universidade Federal de Viçosa - UFV, Departamento de Zootecnia, CEP: 36570-900, Viçosa, MG, Brasil. E-mail: *tallesmaciel@gmail.com; fabyanofonseca@ufv.br; sfacioni@ufv.br*

foi amplamente utilizada para quantificar a abundância de mRNAs, baseando-se na medida da intensidade da fluorescência (medida considerada variável contínua). Por outro lado, RNA-Seq resulta numa medida discreta para inferência da expressão diferencial. Em muitos estudos de RNA-Seq o objetivo central consiste na identificação e quantificação de genes diferencialmente expressos (GDE) entre diferentes tratamentos/condições, visando compreender a base molecular da variação fenotípica.

Devido à crescente utilização desta tecnologia e à inexistência de um método padrão para inferir expressão diferencial (ED), vários métodos foram desenvolvidos com este propósito. Dentre os mais utilizados, tem-se: edgeR (ROBINSON *et al.*, 2010), DESeq (ANDERS e HUBER, 2010) e baySeq (HARDCASTLE e KELLY, 2010) que se baseiam na distribuição de probabilidade Binomial Negativa. Nestes três trabalhos, juntamente com os trabalhos desenvolvidos por Trapnell *et al.* (2012) e Bullard *et al.* (2010) é relatado que esta distribuição é ideal para lidar com problemas de superdispersão encontrados na presença de réplicas biológicas.

Em decorrência do desenvolvimento de diversas metodologias disponíveis para análise de expressão diferencial e na inexistência de uma que seja ótima em todas as situações, a comparação entre estas é importante e tem despertado o interesse dos pesquisadores.

Kvam *et al.* (2012) utilizaram dados simulados baseados em diferentes distribuições de probabilidade e dados reais, para comparar os métodos EdgeR, DESeq, baySeq e o método que emprega o modelo de Poisson de dois estágios (two-stage Poisson model - TSPM). Neste trabalho, os métodos foram comparados quanto à capacidade em discriminar os GDE. Robles *et al.* (2012) avaliaram o impacto do aumento da cobertura gerada pelo sequenciamento, da existência de réplicas biológicas e da utilização de multiplex (sequenciamento de várias amostras no mesmo compartimento da placa de sequenciamento) na detecção de genes diferencialmente expressos. Estes dois trabalhos verificaram ainda a taxa de falsos positivos encontrados pelos métodos avaliados. Soneson e Delorenzi (2013) compararam, por meio de simulação de dados, onze métodos destinados a análise de expressão diferencial. Neste trabalho foi avaliada a concordância e a sobreposição dos resultados.

Na prática, mesmo com estudos demonstrando a importância da existência de repetições biológicas para a inferência da expressão diferencial (KVAM *et al.*, 2012, ROBLES *et al.*, 2012), ainda existem muitos estudos sendo realizados sem repetição ou com poucas repetições (HEITLINGER *et al.*, 2013). Segundo Feng *et al.* (2012), quase 70% das amostras de RNA-Seq de humanos disponíveis no banco de dados de Expressão Gênica Omnibus (Gene Expression Omnibus - GEO) (EDGAR *et al.*, 2002, BARRETT *et al.*, 2013) não tem réplicas biológicas. Este fato deve-se, também, ao elevado custo do sequenciamento (AL SEESI *et al.*, 2014), de delineamentos experimentais inapropriados e da dificuldade em se obter determinados tipos de amostras, como por exemplo, estudos que envolvem a fase pré-natal de suínos, em que as amostras são os fetos.

Apesar da disponibilização de vários pacotes de softwares destinados à identificação de GDE em experimentos de RNA-Seq, poucos trabalhos comparando estas metodologias foram publicados (SEYEDNASROLLAH *et al.*, 2013). Dentre os disponíveis, nenhum avaliou a inferência de expressão diferencial em experimentos sem réplicas biológicas utilizando os pacotes DESeq e baySeq. As metodologias implementadas nos pacotes DESeq e baySeq permitem tal análise e por esta razão foram avaliadas neste trabalho.

Apesar de se basearem na distribuição binomial negativa, elas possuem enfoques estatísticos diferentes (frequentista e bayesiano). Os trabalhos disponíveis também não avaliaram a sobreposição de resultados obtidos por estes dois pacotes considerando cenários com número de repetições variáveis.

Diante do exposto, este trabalho teve por objetivo avaliar, por meio de dados simulados, a influência do número de repetições na identificação de GDE pelas metodologias empregadas nos pacotes DESeq e baySeq bem como avaliar a sobreposição dos GDE encontrados, nos diferentes contrastes estabelecidos, dentro e entre os métodos estudados. Por fim, foi avaliada a taxa de falsos positivos encontrados por estas metodologias.

2 Material e métodos

Foram criados quatro cenários (C1, C2, C3 e C4) mimetizando experimentos reais com o intuito de avaliar as duas metodologias em estudo. Os cenários são compostos por número diferentes de repetições, sendo cada observação/repetição dos cenários composta por 1000 genes. Os cenários foram criados considerando duas condições experimentais (controle e tratado, por exemplo). Cenários denotados por C1, C2, C3, contém, respectivamente, cinco, três e duas repetições. Enquanto, C4, não contém repetição (Tabela 1).

Tabela 1 - Especificações dos cenários analisados

Cenários	nº de repetições	nº de observações*	nº de genes / observações	nº de GDE	nº de não GDE	Genes *up*	Genes *down*
C1	5	5	1000	200	800	100	100
C2	3	3	1000	200	800	100	100
C3	2	2	1000	200	800	100	100
C4	0	1	1000	200	800	100	100

*nº de observação simulada para cada condição experimental; GDE: genes diferencialmente expressos; *up*: upregulados no tratamento em relação ao controle; *down*: dowregulados no tratamento em relação ao controle.

A função *simulateReadCounts* do pacote TCC (an acronym for Tag Count Comparison) (SUN *et al.*, 2013) implementado no Bioconductor (GENTLEMAN *et al.*, 2004) foi utilizada para geração dos dados analisados. Para simulação, considerou-se os 200 primeiros genes DE's e os 800 restantes não DE's. Além disso, considerou-se que os GDE são pelo menos 4 vezes mais expressos que os não DE, ou seja, $\log_{2i}(T_i/C_i) \geq 4$, em que T_i e C_i são, respectivamente, o número de leituras (*reads*) alinhadas no i-ésimo gene. Tal medida é denotada na literatura por *fold-change* (fc) e tem por objetivo quantificar a expressão dos genes. Considerou-se ainda na simulação que, dos 200 GDE, 100 são mais expressos na condição tratado em relação ao controle (*upregulados* – fc>0) e 100 menos expressos na condição tratada em relação ao controle (*downregulados* – fc<0) (Tabela 1).

O procedimento de simulação foi repetido dez vezes, sendo utilizados como resultados, as médias das quantidades de GDE e não DE, além do número de genes sobrepostos, detectados em cada cenário.

Para verificar a influência do número de repetições na detecção de GDE, avaliou-se a expressão diferencial por meio das metodologias implementadas nos pacotes DESeq e baySeq considerando uma taxa de falsa descoberta (*false discovery rate* – FDR) de 5%. Embora o DESeq e baySeq apresentem, respectivamente, abordagens frequentista e bayesiana para realização do teste, as mesmas se baseiam na distribuição Binomial Negativa equação (1).

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

em que, x é número de fracassos até o r-ésimo sucesso, μ é o valor esperado de X e ϕ é o parâmetro de superdispersão.

Conforme descrito por Robinson e Smyth (2007), as curvas FDR são utilizadas para avaliar a proporção esperada de falsos positivos entre todos os testes significativos.

Para avaliar a taxa do Erro Tipo I, considerou-se H_0 verdadeira, isto é, a não existência de expressão gênica. O valor de $\hat{\alpha}$ foi computado pela proporção de rejeição da hipótese nula, com as estatísticas calculadas e valores críticos, obtidos considerando um FRD de 0,05. Para estimar o Poder dos testes, considerou-se H_0 falsa, isto é, genes que apresentam 4 vezes mais expressão que os não DEs.

Em um primeiro momento, os dados foram analisados separadamente por cada método, comparando-se apenas a influência do número de repetições na identificação dos GDE. Em seguida, foi realizada a comparação entre os resultados obtidos por cada método, levando em conta também o número de repetições em cada cenário. Foram construídos diagramas de Venn por meio da biblioteca Venndiagram do R (CHEN e BOUTROS, 2011) com o intuito de avaliar a sobreposição dos GDE encontrados dentro e entre as metodologias analisadas nos quatro contrastes estabelecidos (um para cada cenário entre as diferentes condições).

Visando comparar a performance das metodologias quanto à sensibilidade (ou taxa de verdadeiros positivos), foram construídas as curvas ROC (*Receiver Operating Characteristic*) para cada método e cada cenário. A área sob a curva ROC (*area under the curve* - AUC) foi utilizada como uma medida do desempenho global discriminativo frente aos genes não diferencialmente expressos.

3 Resultados e discussões

3.1 Análise dos dados via DESeq

Conforme pode ser visto, a redução no número de repetições biológicas acarreta a diminuição do poder desta metodologia em identificar os GDE (Tabela 2). Especificamente, o cenário 1 com 5 repetições e o cenário 4 sem repetições apresentaram aproximadamente, maior (58% - 116 GDE) e menor (1% - 2 GDE) percentual de acerto, respectivamente (Tabela 2). Esses resultados corroboram com o estudo realizado por Anders e Huber (2010), que avaliaram a eficiência da metodologia implementada no pacote DESeq para análises de dados de RNA-Seq sem repetições em células neurais,

sendo observado que somente 11% dos genes foram considerados diferencialmente expressos, quando comparados com a análise realizada com duas repetições. Resultados semelhantes corroborando que o aumento do número de réplicas biológicas melhora significativamente o poder de detecção deste método também foram demostrados nos trabalhos de Rapaport *et al.* (2013) e Soneson e Delorenzi (2013). Considerando dados de moscas sem repetições, Anders e Huber (2010) conseguiram identificar 75,09% dos GDE quando comparados ao conjunto de dados que trabalhavam com todas as repetições. Esses resultados evidenciam que a utilização do DESeq para experimentos sem repetições deve ser realizada com cautela visto que sua eficiência está ligada tanto ao número de repetições quanto a variabilidade das amostras.

Tabela 2 - Poder da metodologia DESeq em identificar GDE e taxa de falsos positivos (Erro tipo I); em função do número de repetições

Cenários	nº de repetições	Poder do teste*	nº de GDE encontrados	Erro tipo I**	nº de falsos positivos
C1	5	57,80%	116	0,36%	3
C2	3	40,20%	80	0,84%	7
C3	2	25,10%	50	1,94%	16
C4	0	1,00%	2	6,39%	51

*Capacidade da metodologia em detectar os GDE. **Percentual referente aos falsos positivos dentre os 800 não GDE.

Quanto à taxa de falsos positivos, observou-se que o cenário 4 (sem repetição) apresentou o maior percentual (6,39%). Na prática esse resultado indica que o pesquisador pode levar para análises posteriores genes que não são verdadeiramente diferencialmente expressos, gastando tempo e recursos desnecessariamente.

Observou-se, ao comparar dois cenários por vez, que o maior percentual de GDE comuns foi encontrado ao considerar os cenários 1 e 2 (com 5 e 3 repetições, respectivamente), com 58,75% de concordância. Entre os cenários 1 e 3 (com 5 e 2 repetições, respectivamente), o percentual de concordância foi de 58,0% e entre os cenários 2 e 3 foi de 56,0%. Não houve genes comuns ao considerar o cenário 4 (sem repetição) com nenhum outro (Figura 1a).

Verifica-se ainda que o número de repetições não influencie a identificação dos genes não DE. Estes são basicamente os mesmos tanto com 5 repetições como sem repetição (Figura 1b).

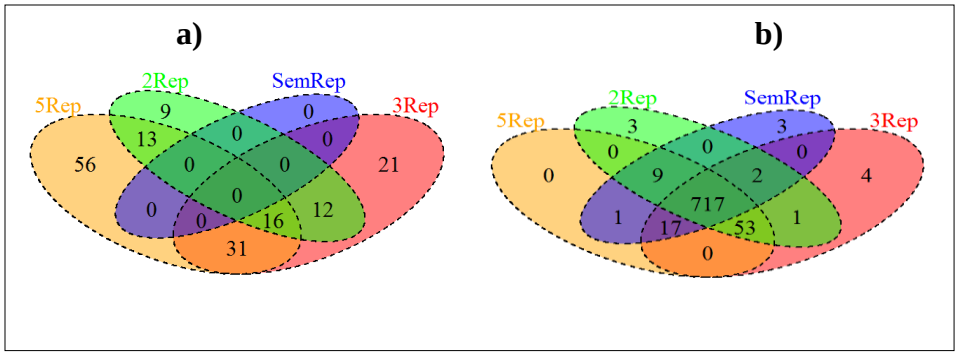

Figura 1 - Diagrama de Venn evidenciando GDE comuns (a) e não comuns (b) aos cenários estabelecidos. 5Rep: C1; 3Rep: C2; 2Rep: C3; SemRep: C4.

3.2 Análise dos dados via baySeq

Assim como ocorreu com a metodologia DESeq, a redução no número de repetições ocasionou uma diminuição do poder do teste implementado no pacote baySeq em identificar os GDE. Os cenários 1 (com 5 repetições) e o 4 (sem repetições) apresentaram respectivamente, o maior (59% - 118 acertos) e menor (0% - nenhum acerto) percentual de acerto (Tabela 3). Esses resultados evidenciam a importância da existência de repetições biológicas em experimentos dessa natureza.

Tabela 3 - Poder da metodologia baySeq em identificar GDE e taxa de falsos positivos (Erro tipo I); em função do número de repetições

Cenários	nº de repetições	Poder do teste*	nº de GDE encontrados	Erro tipo I**	nº de falsos positivos
C1	5	59,00%	118	0,66%	5
C2	3	40,20%	80	0,90%	7
C3	2	23,80%	48	2,08%	17
C4	0	0,00%	0	6,39%	51

*Capacidade da metodologia em detectar os GDE. **Falsos positivos.

À medida que decresce o número de repetições, há um aumento na taxa de falsos positivos, sendo maior no cenário 4. Esses resultados mostram que mesmo modificando o método de análise, quando há decréscimo no número de repetições a taxa de acerto na identificação dos GDE diminui consideravelmente.

O percentual de GDE comuns, ao se comparar dois cenários, foi maior ao considerar os cenários 1 e 2 (com 5 e 3 repetições, respectivamente), com 61,4%; seguido por 58,3% considerando os cenários 1 e 3 (com 5 e 2 repetições, respectivamente) e de 47,9% considerando os cenários 2 e 3 (com 3 e 2 repetições, respectivamente) (Figura 2a).

Para os genes não DE, assim como ocorreu na metodologia DESeq, os resultados são basicamente os mesmos tanto com 5 repetições como sem repetição (Figura 2b).

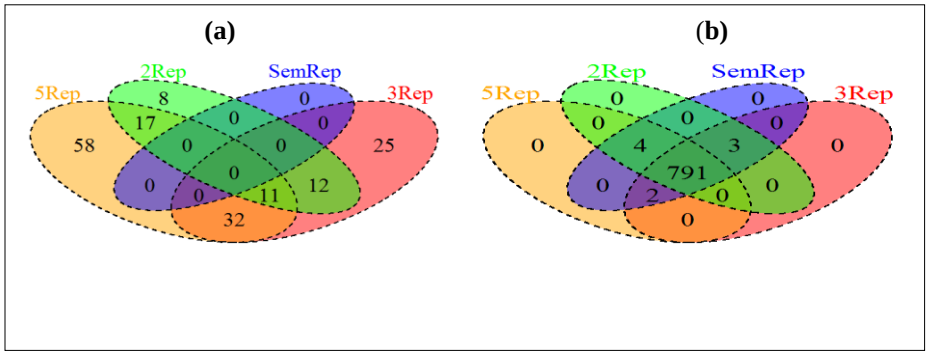

Figura 2 - Diagrama de Venn evidenciando GDE comuns (a) e não comuns (b) aos cenários estabelecidos. 5Rep: C1; 3Rep: C2; 2Rep: C3; SemRep: C4.

3.3 DESeq × baySeq

Seyednasrollah *et al.* (2013) concluíram que o método DESeq é mais conservador. No entanto, neste trabalho, baySeq e DESeq identificaram praticamente o mesmo número de GDE, sendo as taxas de falsos positivos também semelhantes nos dois métodos (Tabelas 2 e 3).

A maioria dos GDE foi idêntica nas duas metodologias, ao considerar o mesmo cenário (Figura 3). Outros trabalhos comparando as metodologias DESeq e baySeq concluíram que nenhum dos métodos é ótimo em todas as circunstâncias e que a escolha do método depende das condições experimentais (RAPAPORT *et al.*, 2013, SONESON e DELORENZI, 2013).

Como esperado, o número de GDE comuns às duas metodologias avaliadas diminui com a redução do número de repetições, sendo 104, 62, 33 e 2 para os cenários 1, 2, 3 e 4, respectivamente. Assim, a escolha de GDE comuns às duas metodologias é uma estratégia que assegura maior confiabilidade na escolha de genes alvos para análises futuras.

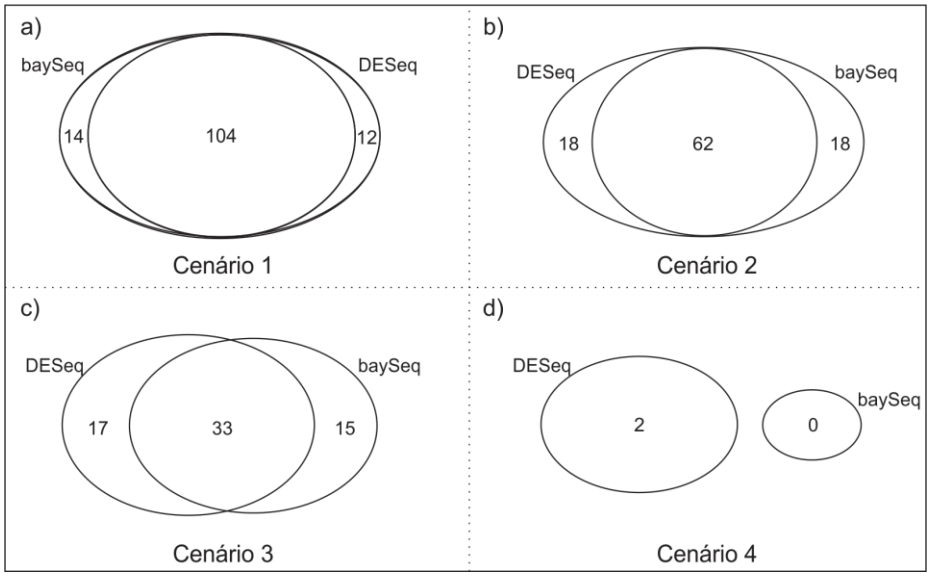

Figura 3 - Sobreposição dos GDE encontrados pelas duas metodologias avaliadas em função do número de repetições. Foi considerada uma FDR menor que 5%.

3.4 Curvas ROC

Nos cenários 1 e 4, o método baySeq apresentou melhor acurácia do que o DESeq para classificar genes verdadeiramente DE, sendo no cenário 4 o desempenho do baySeq (AUC=0,718) significativamente maior do que o do DESeq (AUC=0,502); indicando que para experimentos sem repetição, o baySeq seria o método mais indicado, com uma taxa menor de falsos positivos. Estes resultados são condizentes com a comparação efetuada por Seyednasrollah *et al.* (2013), que concluíram que o DESeq é mais conservador. No entanto, neste mesmo trabalho, os autores afirmam que o pacote baySeq apresentou alta variabilidade. Assim, este resultado deve ser visto com cautela.

No cenário 2 (com 3 repetições) e no cenário 3 (com 2 repetições), a área abaixo da curva (AUC) foi praticamente a mesma. Assim, neste trabalho, pode-se dizer que ambas as metodologias avaliadas apresentaram a mesma acurácia na detecção de falsos positivos, ao considerar 3 e 2 repetições.

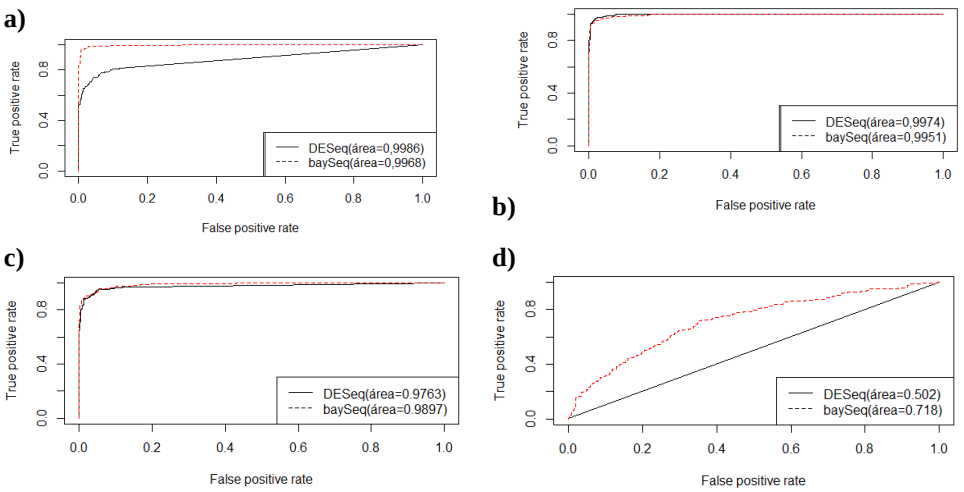

Figura 4 - Curvas ROC para os cenários 1, 2, 3 e 4 (a, b, c, d). O resultado mostra uma pequena vantagem para baySeq na detecção de acurácia nos cenários 1 com 5 repetições e 4 sem repetição. Nos cenários 2 e 3 (com 3 e 2 repetições, respectivamente) a área foi, praticamente, a mesma. Área corresponde à AUC (área abaixo da curva).

Conclusões

Conforme observado, para os dois métodos analisados a redução do número de repetições diminui o poder destas metodologias em detectar GDE. Constatou-se ao comparar dois cenários por vez, para cada metodologia, que um maior percentual de GDE comuns foi encontrado ao considerar os cenários 1 e 2 (com 5 e 3 repetições, respectivamente). Entre os cenários 1 e 3, o percentual de concordância foi menor. O mesmo ocorreu para os cenários 2 e 3. Não houve genes comuns ao considerar o cenário 4 (sem repetição) com nenhum outro.

A maioria dos GDE foi identificada, simultaneamente, pelas duas metodologias, em função do número de repetições.

O método implementado no pacote baySeq apresentou melhor acurácia nos cenários onde havia 5 repetições e no cenário sem repetições. Assim, pode-se dizer que o baySeq apresentou maior sensibilidade, ou seja, maior taxa de verdadeiros positivos e menor taxa de falsos positivos em relação ao DESeq nas condições estudadas.

Apesar deste trabalho auxiliar a escolha entre os métodos baySeq e DESeq para análise de dados de experimentos de RNA-Seq sem e com repetição, cuidados devem ser tomados levando em consideração que outros trabalhos comparativos testaram tais metodologias e concluíram que nenhum dos métodos é ótimo em todas as circunstâncias e que a escolha do método depende das condições experimentais.

Agradecimentos

Os autores agradecem o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Fundação de Amparo à Pesquisa do estado de Minas Gerais (FAPEMIG) pelo auxílio financeiro e aos dois revisores e editores pelos comentários e sugestões.

AMARAL, R. T.; NASCIMENTO, M.; MACIEL, T. E. F.; SILVA, F. F.; NASCIMENTO, A. C. C.; PETERNELLI, L. A.; GUIMARÃES, S. E. F. Number of repetitions in the identification of genes differentially expressed in RNA-Seq studies. *Rev. Bras. Biom.* Lavras, v.36, n.1, p.36-47, 2018.

- *ABSTRACT: This work aimed to evaluate the effect of the number of repetitions, of two important statistical methodologies, BaySeq and DEseq, in the identification of differentially expressed genes (DEG). To carry out the analyses we used four simulated scenarios, whose represents real experiments with two experimental conditions represented for different repetition numbers. TCC package of Bioconductor was used to simulated 1000 genes, which 200 were considered differentially expressed (DE). Initially, the data were analyzed for each method, comparing the influence of the number of repetitions in the identification of DGE. Then, the comparison was made between the results obtained by each method, taking into account the number of repetitions in each scenario. The power to detect DGE was affected negatively due the reducing the number of repetitions. baySeq presented better accuracies for scenarios with 5 and without repetitions. Therefore, baySeq presented higher sensibility, since the rates of true and false positives were, respectively, higher and lower compared to those obtained to DESeq under the evaluated conditions.*

- *KEYWORDS: Transcriptome; simulation; baySeq; DESeq.*

Referências

AL SEESI, S.; TIAGUEU, Y. T.; ZELIKOVSKY, A.; MANDOIU, II. Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates. *BMC Genomics*, v.15, Sup. 8, p.S2, 2014.

ANDERS, S.; HUBER, W. Differential expression analysis for sequence count data. *Genome Biology*, v.11, n.10, p.R106, 2010.

BARRETT, T.; WILHITE, S. E.; LEDOUX, P.; EVANGELISTA, C.; KIM, I. F.; TOMASHEVSKY, M.; MARSHALL, K. A.; PHILLIPPY, K. H.; SHERMAN, P. M.; HOLKO, M.; YEFANOV, A.; LEE, H.; ZHANG, N.; ROBERTSON, C. L.; SEROVA, N.; DAVIS, S.; SOBOLEVA, A. NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Research*, v.41, p.D991-995, 2013.

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

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

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, v.30, n.1, p.207-210, 2002.

FENG, J.; MEYER, C. A.; WANG, Q.; LIU, J. S.; SHIRLEY LIU, X.; ZHANG, Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics*, v.28, n.21, p.2782-2788, 2012.

GENTLEMAN, R. C.; CAREY, V. J.; BATES, D. M.; BOLSTAD, B.; DETTLING, M.; DUDOIT, S.; ELLIS, B.; GAUTIER, L.; GE, Y.; GENTRY, J.; HORNIK, K.; HOTHORN, T.; HUBER, W.; IACUS, S.; IRIZARRY, R.; LEISCH, F.; LI, C.; MAECHLER, M.; ROSSINI, A. J.; SAWITZKI, G.; SMITH, C.; SMYTH, G.; TIERNEY, L.; YANG, J. Y.; ZHANG, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, v.5, n.10, p.R80, 2004.

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

HEITLINGER, E.; BRIDGETT, S.; MONTAZAM, A.; TARASCHEWSKI, H.; BLAXTER, M. The transcriptome of the invasive eel swim bladder nematode parasite Anguillicola crassus. *BMC Genomics*, v.14, p.87, 2013.

KVAM, V. M.; LIU, P.; SI, Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany*, v.99, n.2, p.248-256, 2012.

MOROZOVA, O.; HIRST, M.; MARRA, M. A. Applications of new sequencing technologies for transcriptome analysis. *Annual Review of Genomics and Human Genetics*, v.10, p.135-151, 2009.

RAPAPORT, F.; KHANIN, R.; LIANG, Y.; PIRUN, M.; KREK, 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.

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.

ROBINSON, M. D.; SMYTH, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, v.23, n.21, p.2881-2887, 2007.

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

SEYEDNASROLLAH, F.; LAIHO, A.; ELO, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, v.16, n.1, p.59-70, 2013.

SONESON, C.; DELORENZI, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, v.14, p.91, 2013.

SUN, J.; NISHIYAMA, T.; SHIMIZU, K.; KADOTA, K. TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics*, v.14, p.219, 2013.

TRAPNELL, C.; ROBERTS, A.; GOFF, L.; PERTEA, G.; KIM, D.; KELLEY, D. R.; PIMENTEL, H.; SALZBERG, S. L.; RINN, J. L.; PACHTER, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, v.7, n.3, p.562-578, 2012.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews: Genetics*, v.10, n.1, p.57-63, 2009.

Recebido em 14.04.2016

Aprovado após revisão em 03.05.2017