X Simpósio Brasileiro de Melhoramento Animal
Uberaba, MG – 18 a 23 de agosto de 2013

Análise comparativa de dados de microarranjos submetidos a diferentes métodos de normalização

Bruna Pena Sollero[1], Priscila Grynberg[2], Fabyano Fonseca[3], Glória Regina Franco[2], Simone Eliza Facioni Guimarães[4]

[1]Departamento de Zootecnia – UFMG, Belo Horizonte. e-mail: brunasollero@yahoo.com.br
[2]Departamento de Bioquímica e Imunologia – UFMG, Belo Horizonte. e-mail: priscilag@gmail.com; gfrancoufmg@gmail.com
[3]Departamento de Informática– UFV, Viçosa. e-mail; fabyanofonseca@ufv.br
[4]Departamento de Zootecnia– UFV, Viçosa. e-mail: sfacioni@ufv.br

Resumo: A escolha adequada do método de normalização a ser utilizado em experimentos de microarranjos é crucial para a obtenção de dados de expressões gênicas fidedignos. Por meio de análises de dados brutos de microarranjos representantes do perfil transcricional muscular de suínos, dois métodos diferentes de normalização (*Robust Spline*- RS e *Loess*-L) foram testados com a finalidade de detectar o impacto destes na descoberta de genes diferentemente expressos. Analisando quatro contrastes experimentais, a aplicação do método RS foi capaz de identificar um maior número de genes diferentemente expressos em comparação com o método L (513vs430). Entretanto, observou-se uma alta correlação entre os valores de *fold change* (>0,99) daqueles genes comparados entre análises, independente do método utilizado. De maneira geral, os valores médios de FDR para os grupos de genes comparados apresentaram-se similares, ainda que considerados genes diferentemente expressos somente para um dos métodos de normalização. De acordo com as premissas estatísticas, o número de genes detectados como diferentemente expressos variaram entre os contrastes analisados, conforme aplicação de um ou outro método de normalização. Para obter resultados mais conclusivos, análises adicionais foram propostas.
Palavras–chave: bayesiano, suínos, transcriptoma

A comparative analysis of microarray data set analyzed with different normalization methods

Abstract: Optimal selection of a normalization method to be used in microarray experiments is crucial to generate reliable gene expression data. Using the raw data of a microarray experiment, in which the pig muscle transcriptome profile was investigated, two different normalization methods (Robust Spline- RS e Loess-L) were tested in order to verify their impact on discovering genes differentially expressed. According to analyzes of four experimental contrasts, the application of the RS method was able to identify a greater number of genes differentially expressed in comparison with the L method (513vs430). However, high correlations between fold change values (>0,99) of the genes compared between analyzes were observed independently of the method applied. In general, the averages of FDR values for the group of genes compared among the analyzes were very similar, even if considered genes that presented to be differentially expressed in one but not in the other method. Under statistical assumptions inferred, the numbers of genes identified as differentially expressed varied among the contrasts analyzed according to the application of one or the other normalization method. In order to obtain more conclusive results, additional analyses were proposed.
Keywords: bayesian, pigs, transcriptome

Introdução

Microarranjos, hoje, é considerada uma técnica para análise de expressão gênica amplamente difundida. Neste contexto, existem diversos trabalhos avaliando o impacto de diferentes metodologias e protocolos aplicados em análises destes tipos de dados (Schmid et al., 2010), os quais podem afetar significativamente os resultados e as interpretações do perfil transcricional estudado.

Todas as etapas de experimentos utilizando microarranjos são críticas para a obtenção de dados de expressões gênicas fidedignos. Mais especificamente, a escolha de um adequado método de normalização dos dados é crucial para remover ou minimizar os efeitos sistemáticos que não são constantes entre as amostras no experimento ou que não sejam devido aos fatores objetivamente investigados, pois as análises de descoberta dos genes diferentemente expressos devem representar, coerentemente, os níveis de expressão em uma variedade de condições biológicas.

O presente estudo comparou dois métodos distintos de normalização em um mesmo conjunto de dados de microarranjos previamente analisado por outro protocolo (Sollero et al., 2011), a fim de detectar o impacto destes na descoberta de genes diferentemente expressos relacionados ao processo de desenvolvimento muscular pré-natal de suínos.

Material e Métodos

Este trabalho baseou-se nos resultados da avaliação do perfil transcricional do músculo *Longissimus dorsi* representante de dois grupos genéticos de suínos em duas idades pré-natais por meio de 13 lâminas de microarranjos de oligonucleotídeos (*Swine Protein-Annotated Oligonucleotide Microarray*– www.pigoligoarray.org). Para tal, o *row data* (arquivos *.gpr*) extraído do scanner Axon 118 GenePix® 4000B (*Molecular Devices*) foi analisado por meio do pacote *limma* no *software* R (*R Core Development Team* 2009, versão 2.3.1) de acordo com os seguinte contrastes experimentais, comparando-se: 1) grupos genéticos de suínos aos 40 dias de gestação (C40vsP40), 2) grupos genéticos aos 70 dias de gestação (C70vsP70), 3) idades pré-natais para a raça composta Yorkshire x Landrace (C40vsC70) e 4) idades pré-natais para a raça Piau (P40vsP70).

Após procedimento de correção de *background*, os métodos de normalização dentro de lâminas denominados *Robust Spline* (Smyth & Speed, 2003) e *Loess* (Yang et al., 2002) foram testados, independentemente, no mesmo conjunto de dados. A fim de identificar os genes diferentemente expressos, testes de significância para cada contraste e método de normalização aplicado, foram realizados utilizando-se o método bayesiano empírico de estatística *t* moderada (Smyth, 2004) e um ajuste por meio de FDR (*false discovery rate*), de acordo com Benjamini & Hochberg (1995).

Resultados e Discussão

O primeiro contrate analisado apresentou 221 genes diferentemente expressos (FDR<0,05) nas análises que utilizaram tanto o método *Robust Spline* (RL), como o *Loess* (L). Por outro lado, 110 genes apresentaram-se diferentemente expressos apenas na análise que aplicou o método RS, enquanto somente 13 genes foram diferentemente expressos quando aplicado o método L. Já no segundo contraste, nenhum gene foi detectado como diferentemente expresso, independente dos métodos de normalização testados. Referente ao terceiro contraste analisado, apenas cinco genes foram detectados com FDR<0,05 nas análises em que o primeiro método de normalização foi utilizado e apenas um gene como tal, quando utilizado o método L. Por último, a comparação entre idades pré-natais dentro da raça Piau revelou 148 genes diferentemente expressos em ambos os métodos de normalização aplicados, enquanto 34 foram detectados como diferentemente expressos somente ao testar o RS e outros 47 apenas mediante aplicação do método L. Os resultados referentes a valores de *fold change* e de FDR obtidos pelos contrastes (1 e 4) que exibiram um maior número de genes diferentemente expressos entre os métodos de normalização testados, foram apresentados de forma comparativa (Tabela 1). Pode-se observar uma alta correlação entre os valores de *fold change* (>0,99) daqueles genes comparados, independente de terem sido diferentemente expressos ou não para ambos os métodos. Mais especificamente, os dois métodos de normalização apresentaram-se igualmente efetivos na identificação daqueles genes diferentemente expressos com maiores valores e variâncias de *fold change*. Para aqueles genes detectados como tal, mas exclusivamente para um método, menores valores e variâncias de *fold change* foram observados quando comparados.

O método de regressão não paramétrico *Loess*, é normalmente mais difundido entre os métodos de normalização. Por outro lado, o método *Robust Spline*, que utiliza regressão curva (*spline*) baseando-se em inferência bayesiana, tem se mostrado mais robusto na normalização de alguns dados de microarranjos (Smyth & Speed, 2003). Considerando os quatro contrastes analisados, a aplicação do método RS determinou um maior número de genes diferentemente expressos em comparação com o método L (513vs430), especialmente no contraste entre grupos genéticos aos 40 dias de gestação. Sabe-se que alguns métodos de normalização geram dados com menor variância que outros, resultando em menores valores de *fold change*, entretanto, apresentando *p-values* mais significativos (Schmid et al., 2010). Neste caso, valores médios de FDR para os genes que se apresentaram diferentemente expressos para um método de normalização e não para o outro, foram ligeiramente mais distintos (0,039vs0,071 e 0,069vs0,040) no contraste entre idades dentro do mesmo grupo genético (P40vsP70). Ainda para este

contraste, o método L identificou um maior número de genes (47vs34) diferentemente expressos que o método RS. Entretanto, de maneira geral, os valores médios de FDR para os grupos de genes observados entre as análises (Tabela 1) apresentaram-se próximos, o que também dificulta a definição daquele método mais sensível, e, portanto, recomendável. Somado a isto, similarmente, os dois métodos se apresentaram praticamente incapazes de identificar genes diferentemente expressos em dois dos contrastes analisados (2 e 3).

Uma vez que o método de análise estatística aplicado para a identificação de genes diferentemente expressos foi o mesmo para todas as análises, os resultados, em termos de número de genes diferentemente expressos detectados nos contrastes, apresentaram-se consideravelmente variáveis dependendo do método de normalização aplicado entre os contrastes. Não obstante, os mesmos resultados também contrastam com aqueles primeiramente obtidos (Sollero et al., 2011), o qual utilizou, inclusive, um outro desenho experimental (*loop design*) para proceder as análises.

Tabela 1. Número de genes (N) comparados; estimativas de correlação (Corr) e variância (Var) para valores de *Fold Change* (FC) e média (M) para valores de FDR obtidos em cada contraste (C40vsP40 e P40vsP70) ao testar os métodos de normalização *Robust Spline* (RS) e *Loess* (L).

		FC			*FDR*	
C40vsP40	N	Corr	Var(L)	Var(RS)	M(L)	M(RS)
L&RS	221	0,9985	**5,4871**	**5,3425**	**0,0435**	**0,0399**
L	13	0,9989	**3,7023**	3,2417	**0,0472**	0,0547
RS	110	0,9984	3,4134	**3,4412**	0,0570	**0,0456**
P40vsP70						
L&RS	148	0,997	6,738	6,686	0,0237	0,0250
L	47	0,995	**3,697**	3,333	**0,0396**	0,0712
RS	34	0,999	3,246	**3,259**	0,0692	**0,0407**

Número de genes (N) diferentemente expressos (DE) (FDR<0,05) em ambos os métodos (L&RS); somente na análise pelo método *Loess* (L); somente na análise pelo método *Robust Spline* (RS). Em negrito, valores correspondentes aos métodos capazes de identificar genes diferentemente expressos.

Conclusões

Por meio deste estudo comparativo, foi possível constatar que os dois métodos de normalização de dados de microarranjos testados, apresentaram diferentes resultados com relação ao número de genes detectados como diferentemente expressos. Ainda que o método *Robust Spline*, de maneira geral, tenha se mostrado ligeiramente mais sensível quando comparado com o método *Loess*, os resultados se mostraram variáveis entre os contrastes analisados.

Sugerimos que análises de *cluster* sejam realizadas adicionalmente para comparar as funcionalidades biológicas propostas em cada grupo de genes identificados como diferentemente expressos entre os métodos de normalização testados.

Literatura citada

BENJAMINI, Y. & HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. **Journal of the Royal Statistical Society**, v.57, p.289-300, 1995.

SCHMID, R.; BAUM, P.; ITTRICH C. et al. Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. **BMC Genomics**, v.11, p.1-17, 2010.

SMYTH, G.K. & SPEED, T. Normalization of cDNA microarray data. **Methods**, v.31, p.265-273, 2003.

SMYTH, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. **Statistical Applications in Genetics and Molecular Biology**, v.3, 2004.

SOLLERO, B.P.; GUIMARÃES, S.E.F.; RILINGTON, V.D. et al. Transcriptional profiling during foetal skeletal muscle development of Piau and Yorkshire–Landrace cross-bred pigs. **Animal Genetics**, v.42, p.600-612, 2011.

YANG, Y.H.; DUDOIT S.; LUU P. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. **Nucleic Acids Research,** v. 30, p.1–15, 2002.