Novel algorithm for transcriptome analysis

Peter M. Saama,1 Osman V. Patel,2 Anilkumar Bettegowda,2 James J. Ireland,2 and George W. Smith2

1Genus Plc., Hendersonville, Tennessee; and 2Department of Animal Science and Center for Animal Functional Genomics, Michigan State University, East Lansing, Michigan

Submitted 1 June 2006; accepted in final form 20 September 2006

Saama PM, Patel OV, Bettegowda A, Ireland JJ, Smith GW. Novel algorithm for transcriptome analysis. Physiol Genomics 28: 62–66, 2006. First published September 26, 2006; doi:10.1152/physiolgenomics.00108.2006.—A growing body of evidence implicates the oocyte as a key regulator of ovarian folliculogenesis and early embryonic development. We have screened bovine cDNA microarrays (containing expressed sequence tags representing >15,000 unique genes) with Cy3- and Cy5-labeled cDNA derived from bovine oocyte samples collected at two different stages of meiotic maturation (germinal vesicle vs. metaphase II; n = 3 samples per group). Here, we present a novel data analysis approach that uses all available information from above experiments to obtain and index the transcriptome of bovine oocytes and changes in transcriptome composition in response to meiotic maturation. Signal intensities (Fg) for all housekeeping genes were omitted prior to analysis. A local threshold for gene expression was computed as background intensity (Bg) plus 2 times the standard deviation of background and foreground signals. Within each array, data were normalized by the LOWESS procedure. Subsequently, a two-stage mixed model was fitted to remove systematic variations. In the first stage, the response was the LOWESS normalized Fg with treatment as a fixed effect. In stage 2, the residuals from stage 1 were analyzed in a gene-specific model that included treatment group and spots nested within patch and array. A test for the difference between least squares means for the treatment effect was performed. A false discovery rate (FDR) adjustment on the p values for the difference was carried out. This novel algorithm was compared with approaches that ignore the FDR and the threshold described herein and stark differences obtained.

bovine; microarray; local threshold; mixed model; false discovery rate

ABUNDANT EVIDENCE INDICATES the oocyte is a key regulator of fertility. A developmental program intrinsic to the oocyte controls the overall rate of ovarian follicular development (8), and oocyte-derived transcripts are critical for early embryonic development before initiation of transcription from the embryonic genome (18). However, in general, the composition of the oocyte transcriptome (catalog of genes expressed in female germ cells) and temporal regulation of the majority of oocyte-expressed genes are not well understood. Furthermore, improvements in procedures for in vitro meiotic maturation of oocytes are critical to enhance application of relevant biotechnologies such as in vitro embryo production and nuclear transfer cloning and dependent upon an enhanced knowledge of regulation of meiotic maturation at the transcriptome level.

The merits of microarray-based transcriptome analysis have been established (1, 20). In microarray experiments, optical fluorescence intensity is the main cause for background intensity (19). Background signal can also be attributable to contamination from the hybridization or washing procedures (19). Therefore, it is customary to adjust the observed foreground intensity for the background intensity (21). However, subtracting the signal background from the foreground intensity introduces negative estimates of gene expression for features with a low dynamic range. Consequently, the high local background hinders quantification of mRNA abundance. Model-based approaches for background adjustment have been proposed (3, 7, 11, 13, 22). Most of these methods depend on the proper choice of positive and (or) negative controls. For this reason, a local threshold criterion for detecting differentially expressed genes that does not depend on representative positive and negative controls is appealing.

In transcriptome profiling experiments of this nature, traditional approaches for controlling the error rates in the presence of a large number of comparisons include conservative and liberal control of family-wise error rates (FWER), using procedures such as Bonferroni correction. When multiplicity exists, these procedures can generate false positives and false negatives. Potential sources of multiplicity include comparison of several treatment or dose groups or genes, multiple endpoints, multiple time points, interim analysis, multiple tests of the same hypothesis (e.g., parametric and nonparametric), variable and model selection, and subgroup analysis. The false discovery rate (FDR) provides an alternative quantification of error under a multiplicity of comparisons. Possible outcomes from multiple comparisons are given in Table 1.1

The interpretation of FDR is as follows: suppose that R tests out of m are declared to be significant at an FDR of 0.05, then 5% of these declarations can be expected to be false positives, on average. Let m0 be the number of genes for which the null hypothesis holds with or without underlying differential expression and π0 be the proportion of such genes. An estimate of π0 may be taken as the value that solves the equation (16):

\[ \hat{\pi}_0(\lambda) = \frac{\#(p_i > \lambda)}{m(1 - \lambda)} \]

where \( p_i \) are the observed (i.e., raw) p values, and 0 < \( \lambda < 1 \) is a tuning parameter and was assumed to be 0.09. In multiple testing, the multiplicity-adjusted p value for a particular null
hypothesis being tested is the smallest FWER at which the test may be declared significant. Analogously, the q value (16) is the smallest FDR at which the test may be declared significant:

\[
q = \text{value}(p) = \min_{\alpha \leq p_i} \text{FDR}(\alpha)
\]

We have conducted experiments using a bovine cDNA microarray representing ~15,200 unique genes (17) to partially characterize the bovine oocyte transcriptome and changes in transcriptome composition in bovine oocytes collected at two stages of meiotic maturation [germinal vesicle (GV) and metaphase II (MII)]. In this report, an estimate of \(\pi_0\) from a two-component mixture model (2) was used to obtain an FDR cut-off for the Benjamini and Hochberg (4) step-up procedure. We then compared the number of significant genes with and without FDR adjustments and with and without consideration of the "signal above background" threshold. We anticipated that using FDR adjustments would lead to the inclusion of genes that would otherwise have been ignored by inferences derived from raw p values.

MATERIALS AND METHODS

GV and MII oocyte collection. Ovaries from adult animals were collected at a local abattoir and transported to the laboratory in sterile 0.25 M NaCl. Upon return to the laboratory, ovaries were washed in sterile 0.25 M NaCl. cumulus-oocyte complexes (COCs) were aspirated and selected (those with more than four compact layers of cumulus cells and homogenous cytoplasm), and cumulus cells were denuded as described previously (5). The denuded GV oocytes (3 pools of 20 oocytes) were snap-frozen in 100 \(\mu\)l of lysis solution (RNAqueous Micro Kit; Ambion, Austin, TX) and stored at ~80°C until RNA isolation.

For collection of MII oocytes, GV stage COCs (from adult ovaries; collected as described above) were matured in vitro as described previously (5). Oocytes with expanded cumulus were denuded, selected based on the presence of a single polar body, and processed in groups of 20 (n = 3) as described above.

RNA extraction. Total RNA was extracted from each pool of GV and MII oocytes using the RNAqueous micro kit (Ambion) according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions. RNA was eluted twice from the silica-based microfilter cartridge using a 10-\(\mu\)l volume of prewarmed (75°C) elution solution according to the manufacturer’s instructions.

Total RNA amplification and cDNA microarray analysis. Total RNA (10 \(\mu\)l) from the pools of GV and MII oocytes (n = 3 each) was amplified using the RiboAmp kit [Arcturus, Mountain View, CA, as described previously (14)]. The quality and quantity of the amplified RNA generated were estimated with a UV spectrophotometer (Beckman Instruments, Fullerton, CA) and the Bioanalyzer 2100 RNA 6000 nanochip (Agilent Technologies, Waldbronn, Germany). Microarray experiments were conducted using procedures described previously (14) and a bovine cDNA array containing expressed sequence tags (ESTs) representing ~15,200 unique genes (17). A total of 15 \(\mu\)g of amplified RNA from GV and MII oocytes were used for cDNA synthesis and labeling.

Statistical analysis. A novel algorithm for characterization of the oocyte transcriptome was developed as follows. Firstly, within channel (Cy3 or Cy5), the local “signal above background” threshold for significance was (1):

\[
F_g > B_g + 2 \frac{\Var[B_g]}{2} + \frac{\Var[F_g]}{2}
\]

where \(F_g\) is the foreground signal intensity, \(B_g\) is the background intensity, and \(\Var\) is the variance operator. Subsequently, the \(F_g\) for each feature on the array was evaluated to determine if it was above or below the “signal above background” threshold. In the current study, the mean signal intensity, mean background variance, and mean foreground variance were used to compute the threshold. The raw data were then normalized using a locally weighted scatter plot smoothing (LOESS) model as implemented in the LOESS procedure of the Statistical Analysis System (15). A two-step mixed model (21) approach was used to account for all known sources of variability and to estimate differences between GV and MII. In the first step, we fit the model:

\[
y_{ijk} = \mu + T_i + e_{ijk}
\]

where \(y_{ijk}\) is the LOWESS normalized spot intensity for \(i^{th}\) gene on \(k^{th}\) array; \(\mu\) is the overall mean, \(T\) is the fixed effect of treatment (i = GV, MII), and \(e_{ijk}\) is the random residual error. In the second step, \(\xi_{ijkl}\), the residuals for each gene \(i\) were used to fit the model:

\[
\xi_{ijkl} = P_n + S_{(i)k} + r_{ijkl}
\]

where \(P_n\) is a fixed effect for the printing pin, and \(S\) is a random effect of spot nested within printing pin and array. The FDR was determined first from raw p values corresponding to significance for the test that the least squares means (LSM) for treatment group was equal to zero. In addition, FDR correction was applied to raw p values for contrasts between LSM for the treatment groups. The Benjamini-Hochberg FDR (4) adjustments for the significance of the hypothesis tests were performed using the MULTTEST procedure of SAS (15). A compound symmetry covariance structure was assumed for the spot within printing pin and array. The FDR was determined first from raw p values corresponding to significance for the test that the least squares means (LSM) for treatment group was equal to zero. In addition, FDR correction was applied to raw p values for contrasts between LSM for the treatment groups. The Benjamini-Hochberg FDR (4) adjustments for the significance of the hypothesis tests were performed using the MULTTEST procedure of SAS (15). A two-component mixture model (2) for the distribution of p values was used to estimate the proportion of genes that were differentially expressed under the null hypothesis. Additional details concerning the analytical procedures that were used for FDR correction with some sample SAS source code are shown in the APPENDIX.

Real-time PCR validation of transcriptome analysis. Real-time PCR procedures were utilized to confirm oocyte expression of a subset of genes determined to be components of the bovine oocyte transcriptome based on results obtained from analysis of described microarray data using above algorithm. Approximately 15 genes located either just above the detection threshold or just below the median point of genes above the detection threshold were selected for analysis. Gene name, GenBank accession number, and primer sequences for genes selected are detailed in Table 2. Procedures utilized for real-time PCR analysis were as described previously (5) with cDNA derived from RNA isolated from GV and MII oocytes used as template. Criteria used for confirmation of oocyte expression of individual genes by real-time PCR included obtainment of an amplification profile where threshold was reached and an amplification plateau was obtained within 38 cycles and obtainment of a single peak of predicted Tm following melting curve analysis.

RESULTS AND DISCUSSION

Determining the FDR cut-off. A standard analysis for gene expression data uses an FDR cut-off of 5% that is motivated by the traditional \(p < 0.05\) cut-off for estimating a critical region (10). Arbitrary FDR cut-offs of 10% (9) or higher, e.g., 20% (6), have been reported. Figure 1 shows a probability histogram of \(p\) values from the two-component mixture model. This plot differs from the uniform distribution because some of the null hypotheses that were tested did not hold. Only a few of the \(p\) values correspond to genes that were differentially expressed at
the 5% level of significance. While the primary objective was to characterize the oocyte transcriptome, it is worth noting that a majority of the \( p \) values corresponded to genes where the null hypothesis of no difference between GV and MII holds. Therefore, an estimate of \( \pi_0 \) could be obtained from the mixture model. A plot of the quantiles of the \( q \) values and \( p \) values from the two-component mixture model is shown in Fig. 2. A quantile is defined as the fraction (or percent) of points below a given value. For example, the 0.3 (or 30%) quantile is the point at which 30% of the data fall below and 70% fall above that value. The data points in Fig. 2 fall approximately along a 45-degree line. This feature demonstrates no lack-of-fit for the mixture model. The estimated value for \( \pi_0 \) was 0.52; thus an FDR cut-off of 52% was used.

Transcriptome profiles. A total of 1,140 genes were differentially expressed between GV and MII. However, this paper focuses primarily on the value of using the proposed threshold

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>GenBank Accession</th>
<th>Primer Sequence</th>
<th>Detected (+)</th>
<th>Not Detected (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dead box polypeptide-51</td>
<td>BE723527</td>
<td>Forward: GACTCCCAGCAGAAGTACTCA</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Interleukin 11 receptor, alpha</td>
<td>BE808981</td>
<td>Reverse: CCAACAGGACAGAGGACAGAT</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Nerve growth factor, beta polypeptide</td>
<td>BF076406</td>
<td>Forward: CGAACAGCTGGATATCGATGTTG</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Chromatin modifying protein-4C</td>
<td>BF652672</td>
<td>Reverse: CATTAAAGCGTGAAGACAATATA</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Molybdenum cofactor synthesis-1</td>
<td>BE899807</td>
<td>Forward: CGATCTGACAACAGAATGATC</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Catenin (cadherin-associated protein), alpha-1</td>
<td>BG688529</td>
<td>Reverse: CGAAGGATCCATACAGAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selenium binding protein-1</td>
<td>BE755206</td>
<td>Forward: TGCTCAGGCAAGAGGTTT</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Stomatin-like protein-2</td>
<td>AW660565</td>
<td>Reverse: CTTCGGCAGTCTCGAGATCA</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Cytochrome b-245, alpha polypeptide</td>
<td>AW477423</td>
<td>Forward: GGATCTGACCAGAAGGACTTC</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Real-time PCR confirmation of results of transcriptome analysis

Fig. 1. Histogram and density of the predicted \( p \) values from the 2-component mixture model.

Fig. 2. QQ-plot showing quantiles of the \( q \) values from the 2-component mixture model plotted against the quantiles of the raw \( p \) values.
and FDR in characterizing the transcriptome profile rather than gene to gene comparisons of differential expression. Table 3 shows the number of overlapping and unique genes with transcripts represented in the transcriptome of GV and MII oocytes. A gene had to be above the threshold in all arrays to be considered in this analysis. We found that 88% of the genes were above the threshold, whereas 72% of these genes were significantly different from 0 ($p < 0.05$); 69% were significant at the 5% FDR cut-off; 86% were significant at the 52% FDR cut-off. The huge differences between the number of genes found to be significant with and without the threshold suggest that ignoring the threshold when characterizing a transcriptome could produce very liberal results. The FDR cut-off from the two-component mixture model admits more genes as potential candidates for differential expression. Only one out of 15 genes considered above this threshold was not significant. This goes to show that an arbitrary % FDR cut-off may exclude some differentially expressed genes.

The National Bovine Functional Genomics Consortium cDNA array that we used in the present study was designed several years ago from all known EST collections in The Institute for Genomic Research bovine gene index (17). This may explain the detection resolution of 88%. In contrast, the transcriptome study in Arabidopsis ranged between 40.4% and 54.3% with an average of 50.6% (1). In the Arabidopsis study, probes were designed from the complete genome sequence rather than only from available cDNA or EST collections, as was the case for our study. Furthermore, it is not yet known whether the threshold for significance utilized in present experiments discriminates against detection of low abundance transcripts. However, real-time PCR confirmation of oocyte expression of 93.3% (14/15) of genes selected was obtained in the current study (Table 2). The reasons for the one false positive obtained are not yet known and may be due to actual lack of expression, primer failure, and or discordance between archived and actual sequence of this particular amplicon spotted on the array (17). Regardless, percent real-time confirmation obtained does support the utility of described algorithm for transcriptome analysis. Furthermore, the threshold sensitivity of 88% observed in the current study seems reasonable given that we had partial coverage of the genome on the array. The threshold criterion used in the Arabidopsis transcriptome study is identical to the one we used. However, that study did not consider FDR adjustments. Furthermore, that study did not adjust for sources of variation using a probabilistic model. The FDR adjustments and the mixed model analysis add novelty to previously described approaches for transcriptome analysis.

We used a gene-centric local threshold to detect transcripts with signal above background. As already mentioned, alternative thresholds that use spike-in controls have been proposed (3, 7, 13, 22). Ease of implementation and performance are factors that currently influence which method is routinely used to correct for the background intensity in a given experiment. Recall that the mean signal intensity, mean background variance, and mean foreground variance were used to compute the threshold. The choice of which measure to use in generating the threshold depends entirely on the software used for image analysis.

To conclude, a local threshold that is based on the variation in the foreground and background signal provides a potentially meaningful basis for whole transcriptome analysis. The analytical approach described facilitates identification of factors that can explain the variability in the data and design of microarray experiments that allow for statistical treatment of the variability and estimation of these factors. Specifically, the mixed model methodology adjusts for known sources of variability by standardizing the data to allow for estimation of adjusted means for transcripts. Multiplicity adjustments allow direct control over the percentage of false positives and improve on existing methods with respect to the percentage of false negatives. For this study, an FDR cut-off of 0.52 was considered to be adequate and 93% confirmation of oocyte expression of a subset of genes determined to be components of the oocyte transcriptome based on described computational procedures using above cut-off was obtained. Above cut-off is specific to this study and was derived from the $p$ value distribution. It is postulated that the dynamic nature of the novel algorithm presented herein should augment existing transcriptome analysis pipelines.

### APPENDIX

#### Benjamini and Hochberg Step-up Procedure

The step-up procedure is as follows (4):

Order the raw $p$ values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$

Find $k = \max\{k : p_{(k)} \leq \frac{k}{m} \alpha\}$

If $k$ exists, reject tests attributable to $p_{(1)}, p_{(2)}, \ldots p_{(k)}$

Thus, the adjusted $P$ values are given by:

$$
\hat{P}_{(1)} = P_{(1)}
$$

$$
\hat{P}_{(m-1)} = \min \left\{ \frac{m}{m-1} P_{(m-1)} \right\}
$$

$$
\hat{P}_{(m-2)} = \min \left\{ \frac{m-1}{m-2} \hat{P}_{(m-1)} \right\}
$$

$$
\vdots
$$

$$
\hat{P}_{(1)} = \min \left\{ \frac{m}{m} \hat{P}_{(2)}, mp_{(1)} \right\}
$$
The step-up procedure is readily available in the MULTTEST procedure of SAS. The sample code below assumes that an SAS dataset called pvalues contains the variable raw_p with the p values:

PROC MULTTEST PDATA=pvalues FDR;
RUN;

Mixtures of Betas

Under the null hypothesis, the distribution of p values, for any valid test, is uniform on the unit interval, U[0,1]. Any such distribution can be modeled as a mixture of v + 1 component distributions in which the jth component is a beta distribution, \( \beta(a,b) \), with probability density function,

\[
\beta(a,b) | p = I_{[0,1]}(p) p^{a-1}(1-p)^{b-1} / Beta(a,b)
\]

Assuming independence of gene expression levels across genes, the log-likelihood function for a collection of p values corresponding to m tests is (2):

\[
L_{m+1}(p; a, b) = \sum_{i=1}^{m} \ln[\pi_0 \beta(1,1)(p_i) + \sum_{j=1}^{v} \pi_j \beta(a_j, b_j)(p_i)]
\]

where \( p_i \) is the raw p value of the ith test, \( \phi_i \) is the probability that a randomly chosen test from the collection of tests is for a gene for which there is no population difference in gene expression, \( \psi_j \) is the probability that a randomly chosen test is for a gene from the jth component distribution for which there is a true population difference in gene expression. The maximum likelihood estimates for the parameters \( \psi_j, a_j, \) and \( b_j \), was obtained iteratively using the NLMIXED procedure of SAS. The sample SAS code below assumes that the variable raw_p contains the p values:

PROC NLIMIXED DATA = pvalues;
PARAMETERS pi0 = .05 a=2 b=2;
pi1 = 1 - pi0;
loglikelihood = LOG(pi0 + pi1*PDF('BETA', raw_p, a, b));
MODEL raw_p = GENERAL(loglikelihood);
RUN;

GRANTS

This work was supported by the Rackham Foundation, the Michigan State University Office of the Vice President for Research and Graduate Studies, and the Michigan Agricultural Experiment Station.

REFERENCES