GENE SET TESTING TO CHARACTERIZE MULTIVARIATELY DIFFERENTIALLY EXPRESSED GENES

John R. Stevens

S. Clay Isom

Recommended Citation

GENE SET TESTING TO CHARACTERIZE MULTIVARIATELY DIFFERENTIALLY EXPRESSED GENES

John R. Stevens[1] and S. Clay Isom[2]

[1] Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, UT 84322-3900 USA; [2] Department of Animal, Dairy, and Veterinary Sciences, Utah State University, 4815 Old Main Hill, Logan, UT 84322-4815 USA

Abstract

In a gene expression experiment (using oligo array, RNA-Seq, or other platform), researchers typically seek to characterize differentially expressed genes based on common gene function or pathway involvement. The field of gene set testing provides numerous characterization methods, some of which have proven to be more valid and powerful than others. These existing gene set testing methods focus on experimental designs where there is a single null hypothesis (usually involving association with a continuous or categorical phenotype) for each gene. Increasingly common experimental designs lead to multiple null hypotheses for each gene, and the characterization of these multivariately differentially expressed genes is of great interest. We explore extensions of existing gene set testing methods to achieve this characterization, with application to a RNA-Seq study in livestock cloning.

Keywords: gene expression, gene set testing, gene set enrichment

1 Introduction

A now-common tool in agricultural research is gene expression technology – whether oligo array (Lockhart et al. 1996), next-generation sequencing (Hayden 2009), or some other platform. These technologies allow simultaneous assessment of the expression of all genes in an organism. This simultaneity lends the ability to assess the activity of biological processes (Subramanian et al. 2005). An ongoing collaboration with animal reproduction researchers uses the RNA-Seq platform (Wang et al. 2009) as part of a long-term goal to understand the molecular mechanisms governing embryo viability and survival in order to improve the efficiency of assisted reproductive technologies for livestock.

cell type **C**		ICM			TE		
	IVV	*	**	*	*	*	*
embryo type **T**	IVF	*	*	*	*	*	**
	NT	*	**	*	*	*	*
	PA	**	**	*	**	**	*
gestation days **G**		10	12	14	10	12	14

Table 1: Summary of design of motivating RNA-Seq experiment. Each asterisk represents a replicate.

1.1 Motivating RNA-Seq Experiment

The expression of 41,693 genomic regions (referred to hereafter as "genes") was assessed in each of 31 pig embryo samples. Three controlled factors were of interest – cell type (C), embryo type (T), and gestation days (G). Cell type (C) had two levels – ICM (inner cell mass, or embryonic stem cells) and TE (trophectoderm, or cells that will become the placenta). Embryo type (T) had four levels – IVV (in vivo fertilization), IVF (in vitro fertilization), NT (somatic cell nuclear transfer, or clone), and PA (parthenogenetic activation, or maternal cloning). Gestation days (G) had three levels, representing specific developmental structure stages – 10 (spherical structure), 12 (tubular structure), and 14 (elongated filament structure). The 31 samples represent partial replication of the $2\times4\times3 = 24$ full factorial combinations of these three factors' levels, as summarized in Table 1.

The general objective in this experiment was to identify and characterize genes with expression differences between embryo types across gestation days. More specifically, the objective was to determine which biological processes were significantly more active (or less, or no different) in certain embryo types than others at specific gestation days. Testing biological processes' activity levels with gene expression levels is possible using gene set testing methods.

1.2 Gene Set Testing

The characterization of significantly differentially expressed genes is most commonly done with respect to Gene Ontology (The Gene Ontology Consortium 2000) or KEGG pathway (Ogata et al. 1999) annotation in the appropriate organism. Briefly, these resources summarize knowledge about genes' functions in specific organisms, such as their involvement in specific biological processes in pig. Rather than focusing on individual genes' expression differences between experimental conditions, it is possible to turn attention to these biological

processes (for example) to test whether a given biological process (or the set of genes with annotation to that process) is differentially active between experimental conditions. Thus gene expression technology can be used to test whether biological processes are "behaving" much different in two experimental conditions.

The literature on such "gene set" testing is too vast to review in full here, but the relevant statistical issues and a comparison of existing methods have been reported elsewhere (Fridley et al. 2010). Briefly, the gene set testing methods found to be most powerful (Fridley et al. 2010) are Fisher's p-value combination method (Fisher 1932) and Goeman's global test (Goeman et al. 2004). Section 2.3 below addresses the relevance and use of these findings for the motivating experiment.

2 Methods

2.1 Multivariate Differential Expression

In the motivating experiment (Section 1.1), the questions of greatest interest involve the identification of biological processes whose activity levels respond (up or down) to the effect of the in vitro culture (IVV vs. IVF comparison), the effect of cloning (NT vs. IVF comparison), or the effect of a maternal contribution only (PA vs. IVF comparison), in each cell type (ICM, TE) separately, and at each gestation day (10, 12, 14) separately. At the same time, it is of interest to identify biological processes that show no difference in activity between these embryo types at each gestation day separately. For one hypothetical example, maybe in the trophectoderm (cell type C = TE) a given biological process is more active in IVV than IVF at gestation day 10, but has no activity difference at days 12 and 14. For another hypothetical example, in the inner cell mass (cell type C = ICM) a given biological process may be less active in IVV than IVF at day 10, then more active in IVV than IVF at day 12, but has no activity difference at day 14.

Because there are three possible activity conclusions (higher than, lower than, no different than IVF) at each of three gestation days (10, 12, 14), there are $3 \times 3 \times 3 = 27$ possible relativized profiles (vs. IVF) to which a biological process could be assigned, for each cell type (ICM/TE) and each comparison of interest (IVV/NT/PA vs. IVF). Some of these are visualized in Figure 1, including the two hypothetical scenarios posed in the preceding paragraph. Only three of the twenty-seven possible profiles are represented in Figure 1. The black solid line represents IVF (the "control" in the comparisons of interest), while the red dashed line represents either IVV, NT, or PA (the "treatment" groups). The numeric legend at the left of Figure 1 summarizes the relative differences between the "treatment" group and IVF at the three gestation days. Red higher than black (numeric value 1 in the legend at left) indicates that the activity level of the biological process is greater in the "treatment" group than in IVF at that gestation day. Red lower than black (numeric value -1) indicates a lower activity level than in IVF. Red at the same level of black (numeric value 0) indicates no activity difference compared to IVF. Each biological process is to be assigned to one of the 27 such profiles in each cell type (ICM/TE) and each comparison of interest (IVV/NT/PA vs.
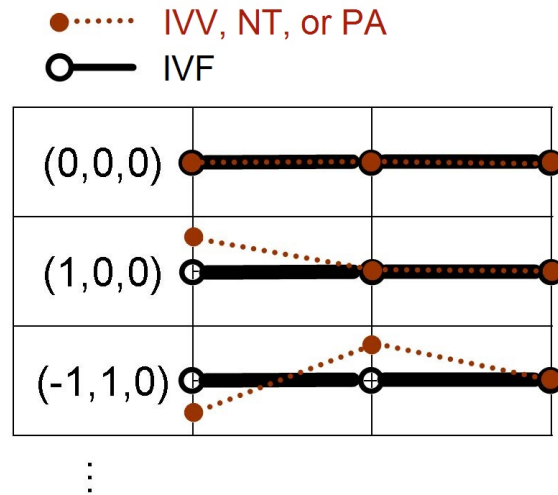
Figure 1: Visualization of select relativized profiles of biological process activity (in "treatment" vs. IVF) across three gestation days.

IVF). This visualization in Figure 1 helps reinforce and reframe the experimental objective (Section 1.1) – to assign each biological process to one of these 27 profiles, for each cell type (ICM/TE) and each comparison of interest (IVV/NT/PA vs. IVF).

A traditional test for differential expression looks at a single significance test for each gene or gene set. However, the experimental objective of the motivating example here leads to multiple simultaneous tests for each gene (and subsequently, gene set). From this arises the notion of what we term "multivariate differential expression" – the idea that multiple comparisons could be of simultaneous interest for a gene or gene set, and a gene or gene set could be differentially expressed in some comparisons (like less active in IVV than IVF at day 10, and more active in IVV than IVF at day 12) but not in others (like no difference between IVV and IVF at day 14; see the bottom profile in Figure 1). A specific approach for gene set testing in situations allowing multivariate differential expression is presented in Section 2.3.

2.2   Generalized Linear Model

The individual components of multivariate differential expression (i.e., the tests of IVV, NT, or PA vs. IVF at each gestation day, in each cell type) can be obtained using contrasts defined for an appropriate per-gene generalized linear model. The statistical issues related to such models for RNA-Seq data have been summarized previously (Auer and Doerge 2010).

Here, let N_l be the total count of all mapped fragments in lane (or replicate) l of the RNA-Seq experiment. Also let Y_{ijkl} be the raw (unscaled) fragment count for the gene in replicate l of gestation day (G) k, for embryo type (T) j, in cell type (C) i. Then $Y_{ijkl} \sim Poisson(N_l p_{ijkl})$, where p_{ijkl} is the probability that any given fragment's sequence

maps to the gene in lane l. The log link in a generalized linear model then yields

$$\log\left(E[Y_{ijkl}]\right) \;=\; \log\left(N_l\right) + \mu + T_j + G_k + TG_{jk}. \tag{1}$$

This model was fit for each gene individually, and for the two cell types separately. The analysis was done using a custom routine in R (R Development Core Team 2012). There was no evidence of widespread overdispersion (results not shown), so the Poisson model was deemed appropriate.

Based on the model in Equation 1, eighteen $(3 \times 3 \times 2)$ contrasts were constructed to test the multiple comparisons of interest for each gene – three "treatments" (IVV, NT, PA) vs. IVF at each of three gestation days (10, 12, 14), in two cell types (ICM, TE) separately. Because the experimental objective here requires a declaration of direction (up or down) relative to IVF, the p-values from each contrast were converted to one-sided p-values using the relative expression of the "treatment" (IVV, NT, or PA) and IVF samples at the appropriate gestation day. This resulted in 18 one-sided p-values for each of 41,693 genes, where for each of the 18, the null hypothesis was of the form "same expression in treatment group as in IVF at this gestation day in this cell type" and the alternative hypothesis was of the form "*greater* expression in treatment group than in IVF at this gestation day in this cell type."

2.3   P-Value Combination Methods for Gene Sets

As mentioned in Section 1.2, the gene set testing methods previously found to be most powerful (Fridley et al. 2010) are Fisher's p-value combination method (Fisher 1932) and Goeman's global test (Goeman et al. 2004). However, Goeman's global test does not lend itself to a multivariate sense of differential expression, as it models how a *single* "clinical outcome" (such as treatment group or continuous phenotype) can be modeled (or predicted) from the expression values of all genes in a given gene set. On the other hand, it is fairly straightforward to use p-value combination methods to obtain a single p-value for each gene set (based on the p-values for all of the genes in the set), for each individual component of multivariate differential expression, such as IVV vs. IVF at day 10 in ICM.

2.3.1   Fisher's and Stouffer's Methods

While Fisher's p-value combination method was found previously to be most powerful (Fridley et al. 2010), it seems that it may be most powerful for a less meaningful alternative hypothesis. In this method as applied to gene set testing (and only considering a single component or test for differential expression, such as IVV vs. IVF at day 10 in ICM), there is a single one-sided p-value (resulting from the appropriate contrast) for each gene annotated to a particular biological process. The null hypothesis is the same for all genes in the gene set (such as $IVV = IVF$ at day 10 in ICM), and so is the alternative hypothesis (such as $IVV > IVF$ at day 10 in ICM). The need for one-sided p-values with common alternatives in p-value combination methods was originally emphasized by Fisher (Fisher 1948), to preserve interpretation of the combined result. The p-values (p_v's) for the genes

in the gene set are combined to obtain a single p-value (being the tail probability of the test statistic $\sum -2\log p_v$, following a chi-square distribution (Fisher 1932)) for the gene set. The global null hypothesis for this combined p-value is the same as the common null for the individual genes' p-values (such as $IVV = IVF$ at day 10 in ICM), and the global alternative hypothesis for this combined p-value is that the common alternative hypothesis (such as $IVV > IVF$ at day 10 in ICM) is true *for at least one gene in the gene set.*

In order to understand why this alternative is not necessarily the most meaningful, it is necessary to note what is actually meant by gene set membership annotation. A gene is annotated to a biological process only when the gene's product "contributes to" the biological process (Hill et al. 2008). (Consequently, there is no annotation if a gene's product impedes or inhibits the biological process.) Then for a biological process to proceed, it is not necessarily sufficient for "at least one" of the contributing genes to be active. In fact, lower activity by any of the genes annotated to a biological process will "disturb" the biological process (Hill et al. 2008). Thus a more meaningful alternative in gene set testing would be that there is a consensus of activity among gene set members – for example, that there is "collective support" (Rice 1990) that the genes annotated to the biological process are more active in IVV than in IVF at day 10 in ICM.

The unsuitability of Fisher's method in cases where consensus is the desired alternative has been shown previously (Rice 1990), as has been the superiority of Stouffer's method (Stouffer et al. 1949) in such cases (Whitlock 2005). Briefly, Stouffer's method converts one-sided p-values to standard normal variates, the weighted sum of which is the test statistic following a standard normal distribution, and whose tail probability is the combined p-value. The global null hypothesis for this combined p-value is the same as the common null for the individual genes' p-values (such as $IVV = IVF$ at day 10 in ICM), and the global alternative hypothesis for this combined p-value is that there is collective support for the alternative hypothesis (such as $IVV > IVF$ at day 10 in ICM) among the genes in the gene set.

2.3.2 Sidedness and Interpretability

Among the biological processes (gene sets) considered in the motivating example, approximately half exhibited a bimodal distribution of p-values. This is summarized for the ICM cell type comparisons in Figure 2. For each gene set, the p-values for each of the 9 contrasts (3 embryo types vs. IVF, at each of 3 gestation days) were treated as following a $Beta(\alpha, \beta)$ distribution, and method-of-moments estimates of α and β were obtained. Figure 2 provides a scatterplot of the resulting estimates, with points colored according to local density (Gentleman and Biocore 2012). Reference lines are added to summarize (based on α and β parameter values) the basic shapes of distributions possible, along with hand-written percentages reporting how many of the plotted points (gene sets) fell to each shape. For example, if α and β are less than 1, and $\alpha > \beta$, the distribution of p-values is essentially bimodal, with the potential for the right mode (near 1) to be higher than the left mode (near 0); approximately 36 percent of the gene sets considered had p-value distributions that fell in this shape category. Note that with the alternative hypothesis that $IVV > IVF$ at day 10,
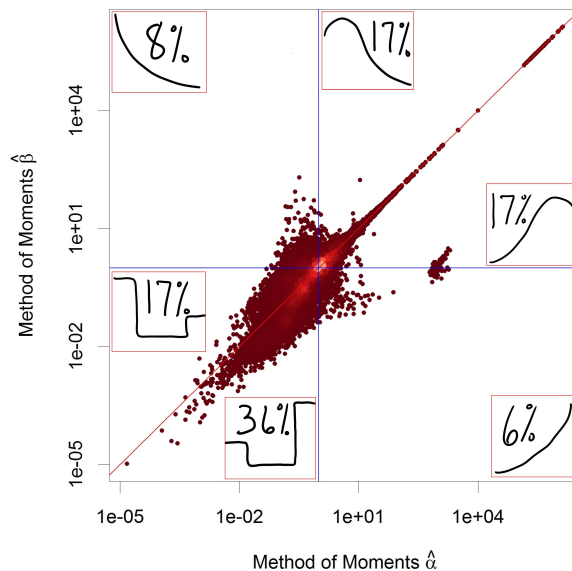
Figure 2: Summary of shapes of distributions of p-values for genes in 5,077 biological processes, based on tests of 9 contrasts in ICM. Approximately half of the biological processes exhibited at least roughly bimodal-shaped distributions.

a gene set with such bimodality (and tall peak near one) could, due to the use of one-sided p-values here, actually be considered as providing evidence of consensus that $IVV < IVF$ at day 10, particularly if the peak near one were much higher than the peak near zero.

In addition to yielding a more meaningful alternative for gene set testing (Section 2.3.1), Stouffer's method is also less sensitive to extreme individual p-values than Fisher's method and has the benefit of symmetry (Whitlock 2005). That is, switching the direction of the alternative (such as $IVV < IVF$ rather than $IVV > IVF$) of the individual genes' tests will result in the complement for Stouffer's p-value (yielding one minus the combined p-value with the original alternative), which will usually not be the case using Fisher's method. (Note that symmetry here refers not to the distribution of p-values, but to the behavior of the combination method when reversing the direction specified by the alternative hypothesis.)

The value of symmetry and consensus in p-value combination methods, particularly for sets of genes with bimodal p-value distributions, can be visualized with a small simulation. 1000 sets of p-values were simulated, each of size 100, from $Beta(\alpha, \beta)$ distributions, at each of various values of α and β. For purposes of discussion, assume that the null hypothesis for each p-value is $IVV = IVF$ at day 10, and the alternative is $IVV > IVF$ at day 10. To emphasize the notion of consensus here, only α and β values less than 1 were used, producing bimodal distributions of p-values. At each ordered parameter pair (α, β), each simulated set of 100 p-values was combined using both Fisher's and Stouffer's methods, and these combined p-values were averaged across the 1000 simulations for each method separately. These are summarized by the contour plots in Figure 3. Fisher's method yields smaller combined p-values than Stouffer's method does for these bimodal distributions. This could
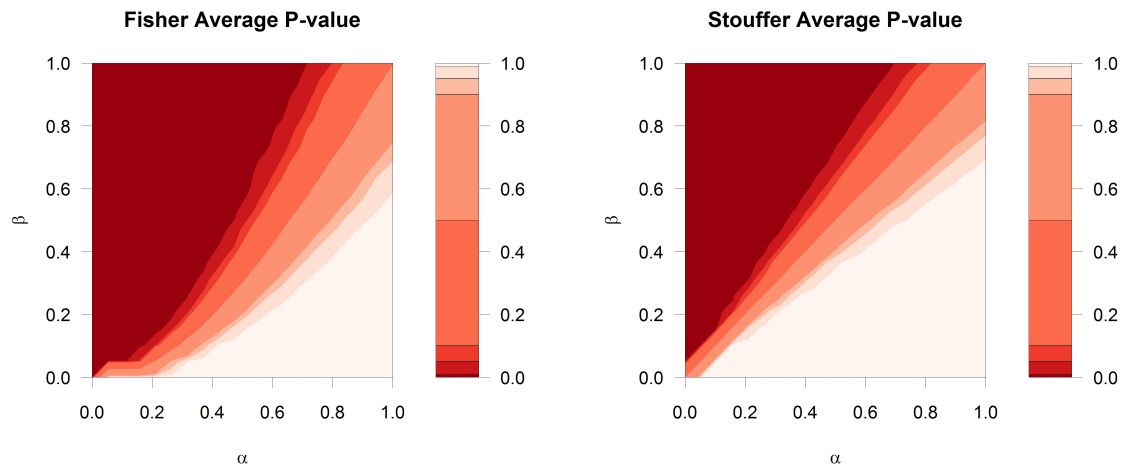
Figure 3: Contour plots of average combined p-values under Fisher's and Stouffer's methods, under various $Beta(\alpha, \beta)$ distribution shapes, in a small simulation study.

result in a nonsense conclusion by Fisher's method, as in the bimodal distribution sketched at the bottom middle of Figure 2 – Stouffer's method would report a high p-value (near 1), indicating evidence that $IVV < IVF$, while Fisher's method could report either a middle-ground p-value (closer to 0.5) or (with a small-enough p-value from a single gene in the set) a low p-value (near 0), indicating evidence that $IVV > IVF$. This (Fisher's method conclusion) would be nonsense when the shape of the distribution of p-values clearly indicates collective support that $IVV < IVF$ (as Stouffer's method would conclude, reflecting the consensus peak of p-values near 1).

In summary, while Fisher's method may be more powerful (Fridley et al. 2010) for a broad alternative (Rice 1990), Stouffer's method is more powerful for a more interpretable (Whitlock 2005) alternative with one-sided p-values.

2.4   Implementation

There are three domains in the Gene Ontology annotation system (The Gene Ontology Consortium 2000) – biological process (BP), molecular function (MF), and cellular component (CC). Based on the objectives of the motivating experiment, attention here is restricted to biological processes. To focus attention on the most meaningful biological processes (that are neither too general nor too specific to be useful), a filter is applied (Song and Black 2008). Specifically, only the 49 percent of biological processes with between 5 and 1000 genes annotated thereto were considered.

For each of the resulting 5,077 biological processes considered, p-values of their annotated genes were combined using Stouffer's method, for each of the 18 contrasts constructed previously (Section 2.2). Thus each of the 5,077 biological processes has 18 Stouffer-combined p-values, one for each of the $3 \times 3 \times 2 = 18$ contrasts constructed previously, corresponding
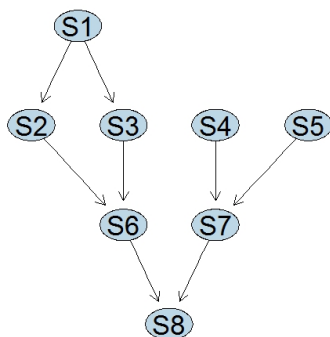
Figure 4: Sample visualization of 8 hypothetical gene sets (biological process terms) in a directed acyclic graph. The arrows indicate nesting or subsetting. For example, set $S1$ is a subset (or special case of biological process) of both $S2$ and $S3$. It is possible that non-subsets (such as $S2$ and $S4$ here) are not disjoint. This potential for nesting and non-disjointness results in possible dependencies among the [combined] p-values for each set.

to tests of three "treatments" (IVV, NT, PA) vs. IVF at each of three gestation days (10, 12, 14), in two cell types (ICM, TE) separately. Using these combined p-values, each biological process can be assigned to one of the 27 relativized profiles (vs. IVF) discussed previously in Figure 1, for each "treatment" (IVV, NT, PA) and each cell type.

Until this point in the analysis, there has been no need for adjustment for multiple comparisons because the p-value combination methods require "raw" (unadjusted) p-values. But in order to assign each biological process to a relativized profile, a statistical decision must be made at each gestation day, based on the appropriate p-value. The strategy employed here was to adjust the 5,077 biological process's p-values within each of the 18 contrasts, to control the false discovery rate (Benjamini and Hochberg 1995). However, the simple Benjamini-Hochberg adjustment (Benjamini and Hochberg 1995) assumes independence of p-values and so would be inappropriate here, as the biological process gene sets are nested and not disjoint (Song and Black 2008), as visualized (Gentry et al. 2011) in Figure 4. Instead, the Benjamini-Yekutieli adjustment (Benjamini and Yekutieli 2001) was applied, as it provides control of the false discovery rate under dependency. The filter mentioned above helped reduce the severity of this adjustment. Because this adjustment is designed for two-sided p-values, the (Stouffer-combined) one-sided p-values for each gene set here were first converted to two-sided p-values, and then converted back to one-sided p-values following adjustment. The false discovery rate was controlled at 0.05 within each of the 18 combinations of embryo type "treatment" (IVV, NT, PA), gestation day (10, 12, 14), and cell type (ICM, TE).

Profile at Day			ICM			TE		
10	12	14	IVV	NT	PA	IVV	NT	PA
0	0	0	4,512	2,969	4,196	2,467	3,037	3,246
0	0	-1	14	1		2,020	1,973	439
0	-1	0	1	2,055		1		566
0	1	0	1	32	881	62	9	3
0	-1	-1		3			1	578
0	1	-1				505	20	
0	0	1	397	14		15	33	13
1	-1	-1						170
-1	0	1	125					
1	-1	0						35
-1	0	0	26					1
1	0	0					3	22
0	-1	1	1	3		5		
1	0	-1						4
0	1	1				2		
-1	0	-1					1	

Table 2: Counts by relativized profile, in "treatment" comparisons vs. IVF, with FDR controlled at 0.05 within each of the 18 combinations of embryo type "treatment" (IVV, NT, PA), gestation day (10, 12, 14), and cell type (ICM, TE). The largest counts for each "treatment" group (column) are highlighted.

3   Results

Each of the 5,077 biological processes was assigned to one of 27 relativized profiles (vs. IVF; see Figure 1) for each "treatment" (IVV, NT, PA) and each cell type, based on its Benjamini-Yekutieli-adjusted Stouffer-combined p-value at each of the three gestation days. Table 3 summarizes the results. The relativized profiles are summarized by a numeric legend at the left of Table 3, indicating whether biological process activity in the "treatment" (IVV, NT, PA) was significantly higher (numeric value 1), significantly lower (numeric value -1), or not significantly different (numeric value 0) than in IVF at each gestation day (10, 12, 14). All relativized profiles with at least one biological process thus assigned are reported in Table 3.

The largest counts for each column in Table 3 are highlighted, and can be used to make some general biological conclusions. For example, in the ICM cell type, there are 2,055 biological processes that are significantly less active in NT than IVF embryo types at gestation day 12, but that show no significant activity differences at days 10 and 14; this is the (0,-1,0) profile. No other profile is so well represented in this column, other than the (0,0,0) profile (showing no significant differences at any gestation day). This suggests that biological process differences in the ICM between NT and IVF embryos occur almost exclusively at gestation day 12, when many biological processes are less active in NT.

As another example, consider the far right column of Table 3. The largest profile counts there show that in the TE cell type, many biological processes are less active in PA than IVF at gestation days 12 (566 processes), 14 (439 processes), or both (578 processes). These are the predominant differences in biological process activity between PA and IVF in the TE cell type.

All of these results are the subject of ongoing interpretation and discussion with animal reproduction collaborators. It may be worthwhile to note in passing that many of the most significant differences vs. IVF for the other embryo types involve biological processes related to "development" – including muscle, skeletal, and brain. It is not surprising that the highest profile counts (Table 3) in ICM and TE cell types do not agree, and this result in fact supports previous observations that biological process activity is fundamentally different in these cell types.

4   Summary

In this manuscript we used a motivating example to build a statistical framework to characterize genes that are differentially expressed (or not) in multiple comparisons of interest, which we term "multivariately differentially expressed." The statistical framework capitalizes on the appropriateness of Stouffer's p-value combination method for such cases. In this motivating example, the characterization was in terms of biological processes from the Gene Ontology annotation system (The Gene Ontology Consortium 2000), and the multiple comparisons of interest corresponded to multiple gestation days. However, the framework is generalizable beyond this motivating example and can characterize in terms of additional annotations (such as molecular function or cellular component) from the Gene Ontology system or KEGG pathways (Ogata et al. 1999).

Future work will include at least two components. First, the Benjamini-Yekutieli adjustment (Benjamini and Yekutieli 2001) utilized in Section 2.4 is a fairly severe adjustment. While this severity is related to the allowance for dependence by this method, we believe it may be mitigated by simultaneously accounting for the known nested structure among Gene Ontology terms. Thus one component of our future work will include adapting this adjustment method to control a meaningful error rate while allowing for dependence within a known structure such as Figure 4. The second component of our future work will involve automating our procedure to characterize multivariately differentially expressed genes, such as in a package for Bioconductor (Gentleman et al. 2004). Briefly, given a matrix of p-values (with rows for genes and columns for specific multiple comparisons of interest) and gene set definitions (such as in a list object in R), we could efficiently create a matrix of Stouffer's combined p-values (with rows for gene sets and columns for the specific multiple comparisons of interest) that have been adjusted for [dependence-allowed] multiple hypothesis testing, and then assign each gene set to various possible profiles across the specific comparisons of interest (as in Table 3). We anticipate using this automated procedure to efficiently characterize multivariately differentially expressed genes in studies conducted by other researchers in animal reproduction and other fields, including some current collaborators.

Acknowledgments

References

Auer, P. L. and R. W. Doerge (2010). Statistical design and analysis of RNA sequencing data. *Genetics 185*, 405–416.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57*, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*(4), 1165–1188.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1948). Combining independent tests of significance. *The American Statistician 2*(5), 30.

Fridley, B. L., G. D. Jenkins, and J. M. Biernacka (2010). Self-contained gene set analysis of expression data: An evaluation of existing and novel methods. *PLoS ONE 5*(9), e12693.

Gentleman, R. and Biocore (2012). *geneplotter: Graphics related functions for Bioconductor*. R package version 1.32.1.

Gentleman, R. C., V. J. Carey, D. M. Bates, and others (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology 5*, R80.

Gentry, J., L. Long, R. Gentleman, Seth, F. Hahne, D. Sarkar, and K. Hansen (2011). *Rgraphviz: Provides plotting capabilities for R graph objects*. R package version 1.32.0.

Goeman, J. J., S. A. van de Geer, F. de Kort, and H. C. van Houwelingen (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics 20*, 93–99.

Hayden, E. C. (2009). Genome sequencing: the third generation. *Nature 457*, 769.

Hill, D. P., B. Smith, M. S. McAndrews-Hill, and J. A. Blake (2008). Gene ontology annotations: what they mean and where they come from. *BMC Bioinformatics 9(Suppl 5)*, S2.

Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology 14*.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research 27*(1), 29–34.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rice, W. R. (1990). A consensus combined p-value test and the family-wide significance of component tests. *Biometrics 46*(2), 303–308.

Song, S. and M. A. Black (2008). Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics 9*, 502.

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. J. Williams (1949). *The American Soldier, Vol. 1: Adjustment during Army Life*. Princeton University Press, Princeton.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS 102*(43), 15545–15550.

The Gene Ontology Consortium (2000). Gene ontology: tool for unification of biology. *Nature Genetics 25*, 25–29.

Wang, Z., M. Gerstein, and M. Snyder (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics 10*, 57–63.

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted z-method is superior to Fisher's approach. *Journal of Evolutionary Biology 18*, 1368–1373.