BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 11 2007, pages 1356–1362
doi:10.1093/bioinformatics/btm116

Gene expression

Extending the pathway analysis framework with a test for transcriptional variance implicates novel pathway modulation during myogenic differentiation

Daniel M. Kemp[1], N. R. Nirmala[2] and Joseph D. Szustakowski[2,*]

[1]Diabetes and Metabolism Disease Area and [2]Genome and Proteome Sciences, Novartis Institutes for BioMedical Research, 250 Massachusetts Avenue, Cambridge, MA 02139, USA

Received on October 30, 2006; revised on March 7, 2007; accepted on March 16, 2007

Advance Access publication March 28, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: We describe an extension of the pathway-based enrichment approach for analyzing microarray data via a robust test for transcriptional variance. The use of a variance test is intended to identify additional patterns of transcriptional regulation in which many genes in a pathway are up- and down-regulated. Such patterns may be indicative of the reciprocal regulation of pathway activators and inhibitors or of the differential regulation of separate biological sub-processes and should extend the number of detectable patterns of transcriptional modulation.

Results: We validated this new statistical approach on a microarray experiment that captures the temporal transcriptional profile of muscle differentiation in mouse C2C12 cells. Comparisons of the transcriptional state of myoblasts and differentiated myotubes via a robust variance test implicated several novel pathways in muscle cell differentiation previously overlooked by a standard enrichment analysis. Specifically, pathways involved in cell structure, calcium-mediated signaling and muscle-specific signaling were identified as differentially modulated based on their increased transcriptional variance. These biologically relevant results validate this approach and demonstrate the flexible nature of pathway-based methods of data analysis.

Availability: The software is available as Supplementary Material.

Contact: joseph.szustakowski@novartis.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Microarray experiments have quickly become staples of twenty-first-century biology. Nevertheless, the analysis of these experiments and extraction of novel biological insights from their results remain an open problem. One of the most common questions biologists seek to address via microarray experiments is 'what pathways or processes are modulated in my system?' Recently several methods have been developed that make use of the currently available repositories of biological pathways to analyze microarray and other high-throughput experiments (for two excellent recent reviews, see Curtis *et al*., 2005; Dopazo, 2006). Although details vary by method, these approaches use various statistical tests and *a priori* biological knowledge, typically in the form of pathways, gene sets or ontological categorizations, to extract additional signal from these experiments. Such pathway-based methods identify groups of functionally related genes that behave in a coordinated fashion across multiple conditions. Application of this knowledge can be done via direct projection onto the microarray results (Mootha *et al*., 2003) or by examining lists of differentially regulated genes for enrichment of genes from specific pathways (Beissbarth and Speed, 2004; Doniger *et al*., 2003; Draghici *et al*., 2003; Khatri *et al*., 2004; Tavazoie *et al*., 1999).

Pathway-based analyses enjoy several advantages over more traditional 'one-gene-at-a-time' methods; taken together, the two approaches offer complementary views of an experiment. By leveraging *a priori* knowledge, pathway-based methods yield more interpretable, hypothesis-based results. Moreover, these methods use statistical tests which are capable of detecting weak signals that would otherwise be missed. As an example, let us consider the gene set enrichment analysis (GSEA) described by Mootha *et al*. (2003) and subsequently extended elsewhere (see, for example, Al-Shahrour *et al*., 2005; Goeman *et al*., 2005; Kim and Volsky, 2005; Kong *et al*., 2006; Szustakowski *et al*., 2006; Tian *et al*., 2005; Tomfohr *et al*., 2005; Zahn *et al*., 2006). The GSEA framework examines all data points in an integrative approach that detects consistent differences between genes in a pathway and all other genes. By comparing the genes of interest to an assumed background distribution (*i.e*. all of the other genes on a microarray chip), this approach offers greatly enhanced statistical power. This 'standard' enrichment approach has proven effective at identifying pathways that are coordinately regulated even when the changes in expression at the transcript level are modest (Mootha *et al*., 2003). Recent extensions of this framework have sought to identify gene sets with a variety of desirable properties. Tomfohr *et al*. (2005) apply a singular value decomposition to identify informative 'meta-genes' in the data. Work by Kong *et al*. (2006) applies multivariate

*To whom correspondence should be addressed.

statistical tests to identify differentially regulated gene sets. Another promising approach attempts to correlate patient survival times with gene sets (Goeman *et al*., 2005).

Pathway-based methods share a common insight, that pathways can be modulated by modest (*i.e*. not statistically significant) but consistent transcriptional changes across many of their constituent genes. This insight has provided a sensitive approach that complements the traditional 'one-gene-at-a-time' approach for analyzing biological data. This insight however applies to only one mode of transcriptional regulation and assumes that available pathway data are accurate and conform to this model. Many pathways can be modulated in more complex and subtle ways than coordinated transcriptional up- or down-regulation. As an example, consider the reciprocal regulation of pathway activators and inhibitors. If a particular pathway is activated, it is reasonable to expect to observe transcriptional up-regulation of that pathway's activators and a concomitant down-regulation of its inhibitors. We were therefore motivated to design additional analysis methods that could recognize more complex transcriptional patterns.

Here, we describe the application of a variance test to expand the repertoire of detectable transcriptional responses. Specifically, we have extended the GSEA framework to include the Levene test for homogeneity of variance as modified by Brown and Forsythe (LBF) (Conover *et al*., 1981). The LBF test operates on median-transformed data (see Methods for details) and has been shown to be robust and powerful relative to other variance tests (Conover *et al*., 1981). The LBF test is intended to complement the standard location-based enrichment analysis. By comparing the transcriptional variance of a pathway to the background variance observed in the experiment, we will be able to detect two additional patterns: pathways with increased variance compared to background, and pathways with decreased variance. Pathways with increased variance are likely to contain many genes that are substantially up- and down-regulated. Such a pattern may be indicative of the reciprocal regulation of pathway activators and inhibitors. When the numbers of up- and down-regulated genes are balanced, the overall expression pattern of the gene set would remain neutral, and the gene set would be undetected by the standard enrichment analysis. Increased variance may also be indicative of the differential regulation of distinct subprocesses that are described in larger gene sets. In contrast, pathways with decreased variance do not necessarily suggest a particular form of modulation. Pathways with decreased variance across a number of conditions may be under tight transcriptional regulation and consequently may be detected more often by standard enrichment analyses as well. In this article, we describe a detailed comparison of the LBF test relative to other statistical tests and demonstrate that it can reveal additional biological interpretations missed by other tests.

2 METHODS

GSEA uses microarray experiment measurements for a system under two different conditions as input. For these analyses, we sorted genes based on their relative expression r_i between condition$_1$ and condition$_2$: $r_i = \mu_{i,1}/\mu_{i,2}$, where $\mu_{i,j}$ is the average expression value for gene i under condition j. Typical comparisons would include diseased versus normal and control versus treated samples.

Each available gene set is projected on to the data one-at-a-time. This projection of *a priori* biological knowledge divides the genes into two groups: those genes known to be involved in a specific pathway k (with corresponding expression ratios R_k) and all other genes (with expression ratios $\tilde{R}_k$). The genes not in a specific pathway serve as a background distribution that reflects the overall biological and technical noise in the data. By applying tests to the expression levels of the genes in these two distributions, one can determine if the behavior of genes in a pathway presents a signal that is detectable with some confidence above the background noise inherent in the experiment. More formally, one can test the null hypothesis $S(R_k) = S(\tilde{R}_k)$, where S is a test statistic chosen to capture a transcriptional signal of interest. This partitioning increases the statistical power of any test used. Whereas a 'one-gene-at-a-time' analysis has statistical power limited by the number of available replicates (typically $n \leq 5$), partitioning the data in this fashion effectively integrates the behavior of all of the genes in a pathway (typically $10 < n < 500$) and exploits the large number of total measurements made on each chip as a background distribution (typically $n > 10\,000$). This overall increase in statistical power makes pathway-based analyses both more sensitive and specific than other 'one-gene-at-a-time' methods.

For this article, we applied three separate statistical tests to the partitioned data: the Wilcoxon ranked sum test, the LBF test and the Kolmogorov Smirnov (KS) test. The Wilcoxon test (Siegal, 1956) represents the standard location-based enrichment analysis and is intended to identify those gene sets with coordinate differential expression. The KS test (Siegal, 1956) is an omnibus test that should be sensitive to a host of distributional differences including (but not limited to) location and spread. The KS test is used here as a reference point to estimate the overall difference in expression patterns for each gene set compared to background. The LBF test is used to locate gene sets with unusually high or low variance compared to background and is applied to $\log(r_i)$ values. The LBF test first transforms the data according to $Z_{ij} = |X_{ij} - \text{median}(X_i)|$, where X_{ij} corresponds to the jth data point from the ith sample. A one-way analysis of variance is then applied to the Z_{ij} values to test for homogeneity of variance. We chose the LBF test over other tests because it provides cleaner results with fewer false positives and it maintains information about the relative magnitudes of the variances tested (see Supplementary Material).

Here, 535 gene sets were used in this analysis, culled from several sources: KEGG ($n = 103$) (Kanehisa *et al*., 2006), Celera/Panther ontology ($n = 204$), Celera public pathways ($n = 56$), Jubilant/Pathart ($n = 171$) and expert curation ($n = 1$). (Links to these and other pathway repositories are provided in Supplementary Material.) We applied the false discovery rate (FDR) q-value multiple testing correction (Storey and Tibshirani, 2003a,b) to all P-value outputs by a specific statistical test to account for the large number of gene sets tested against the data.

We analyzed gene transcription microarray data that captured the differential expression during the temporal progression of muscle stem cell differentiation (Szustakowski *et al*., 2006). The mouse C2C12 myoblast cell line served as a model of late stage (terminal) myogenesis, such that induction of these proliferating mononucleate cells caused cell-cycle arrest and cell fusion to form myotubes that physically contracted and displayed characteristic molecular features of skeletal myocytes.

C2C12 mouse skeletal myoblasts were cultured in DMEM high glucose with 10% FBS and 1% penicillin/streptomycin (Gibco), and maintained at 37°C and 5% CO_2. Differentiation of C2C12 myoblasts into myotubes was achieved by culturing cells in media containing reduced serum concentration (2% v/v) for up to 5 days with media changes every 2 days. C2C12 cell RNA was harvested using the RNeasy

Midiprep kit (Qiagen) following the manufacturer's instructions. Cells were cultured in 6-well plates and a time course of differentiation was performed. Induction of differentiation was initiated at time 0, when cells were confluent, by reducing the serum concentration in the wells to 3% v/v. Cells were lysed for RNA preparation at Days − 1, 0, 0.25, 1, 2, 3, 4 and 5 post-differentiation. RNA was analyzed using the Affymetrix mouse whole genome microarray, MOE430 PLUS 2.0. Each time point was performed in triplicate, from independent experiments.

Microarray data were normalized using the RMA package (Irizarry *et al.*, 2003) with default settings. Normalized values were returned to a linear scale via base-2 exponentiation and scaled to a 2% trimmed mean of 150. Probesets with expression values less than 100 on more than 75% of the chips were discarded as low- or non-expressed. Pathway analyses were performed to compare the transcriptional profiles of C2C12 cells at Day − 1 (myoblasts) and Day 5 (myotubes) of this experiment. An implementation of the methods described here are available as Supplementary Material.

3 RESULTS

Here, 95 of the 535 gene sets were deemed significantly modulated by at least one of the three tests at a FDR q-value threshold of 1E − 3 (see Supplementary Table 1 for results for all 535 gene sets). A graphical overview of these gene sets is presented in Figure 1. The Wilcoxon and LBF tests return almost identical number of gene sets ($n = 50$, $n = 48$, respectively), whereas the omnibus KS test appears to capture the most variation in the data ($n = 61$). When considering all three tests, we observe that the LBF test returns the largest number of unique results ($n = 29$) followed by the KS test ($n = 9$) and Wilcoxon test ($n = 4$). The large intersection of Wilcoxon and KS results ($n = 45$, 90% of Wilcoxon results, 74% of KS results) suggests they are largely capturing similar patterns in the data. In contrast, the LBF test exhibits substantially less overlap with the other tests, sharing 18 out of 48 gene sets with the KS test (38%) and 12 of 48 gene sets (25%) with the Wilcoxon test. These results confirm LBF is sensitive to different patterns than the other methods. These trends also hold true for other q-value thresholds (see Supplementary Fig. 1).

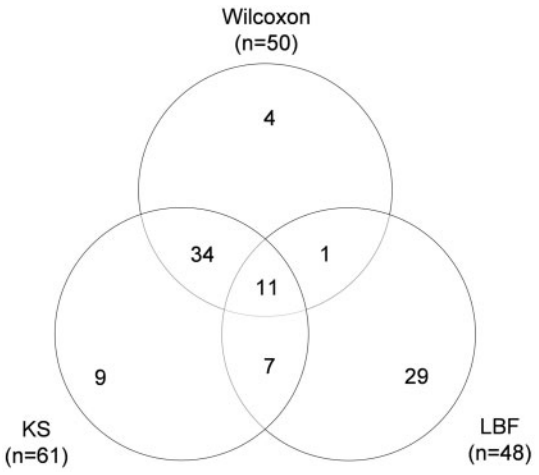

Fig. 1. A Venn diagram depicts the intersection of statistically significant gene sets returned by each of the three statistical tests.

The relative concordance of the test results are reiterated when we take a more granular look at the data. Supplementary Figure 2 presents pairwise scatterplots of the P-values returned by each test for all 535 gene sets used here; Pearson correlation coefficients of log(P-values) are presented in Supplementary Table 2. The LBF and Wilcoxon results show relatively weak correlation (0.45) as do the LBF and KS results (0.51), reinforcing the uniqueness of the LBF results. In contrast, the Wilcoxon and KS tests display good agreement with a correlation of 0.87. Upon inspection, it was determined that the muscle contraction gene set returns highly significant P-values for all three statistical tests and is an outlier for both the LBF test (muscle contraction $P = 1.6\text{E}{-59}$, next best P-value $= 1.4\text{E}{-24}$) and the Wilcoxon test (muscle contraction $P = \sim 0$, next best P-value $= 3\text{E}{-28}$) and may therefore artificially inflate these correlation coefficients. Removing this gene set confirms this observation as the correlation between the LBF and Wilcoxon test is reduced from 0.45 to 0.28, while the other correlations remain relatively unchanged (LBF–KS = 0.51; Wilcoxon–KS = 0.88).

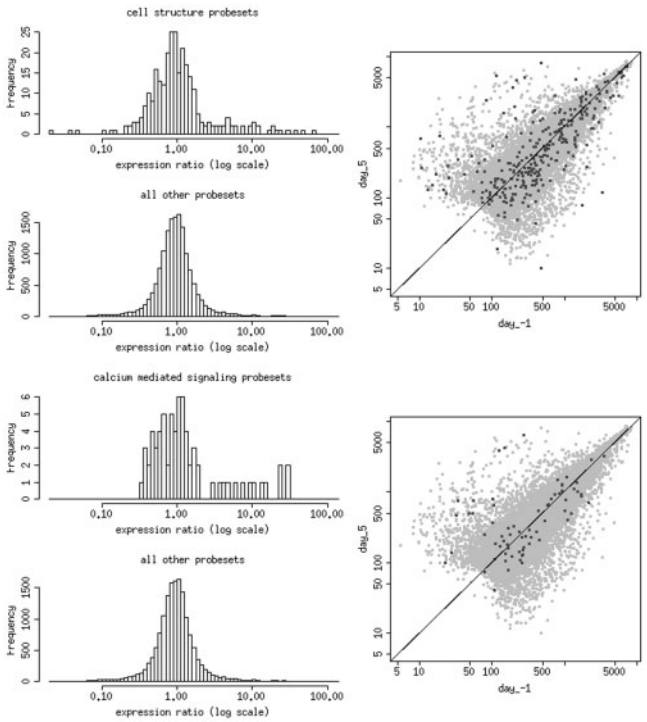

Fig. 2. Example data for two gene sets returned only by the LBF test: 'cell structure' and 'calcium-mediated signaling'. The histograms on the left depict the distribution of expression ratios for probesets within a gene set and all other probesets. The scatterplots on the right indicate the average expression values for all probesets in myotubes at Day 5 versus myoblasts at Day − 1. Black dots mark probesets within the gene set of interest, gray dots indicate all other probesets. Note that in the Wilcoxon test, the P-value returned for these gene sets are not statistically significant because of the approximately equivalent contributions of up- and down-regulated genes.

3.1 Gene sets uniquely identified by LBF

Twenty-nine gene sets were returned as differentially regulated by only the LBF test (Supplementary Table 3). To be useful, the LBF test must return gene sets that are not only unique, but that lead to novel insights or understanding of the underlying biological processes at work during myoblast differentiation. Manual inspection of these 29 gene sets reveals that the LBF results implicate three general processes in myoblast differentiation: cell structure reorganization, Ca^{++}-dependent signaling and muscle-related signaling. These gene sets tend to contain balanced numbers of up- and down-regulated transcripts and are therefore not detected using a conventional enrichment analysis approach with the Wilcoxon test (see Fig. 2 for examples). Taken together, these results validate the viability of variance tests for garnering additional, biologically relevant insights from microarray experiments. A closer examination of these gene sets is presented below.

3.1.1 Cell structure During differentiation of skeletal myoblasts, the morphology and structure of the cell is profoundly altered, such that mononucleated cells fuse to form multinucleated myotubes containing as many as 20 nuclei per cell. This complete reorganization of the cell's structural framework leads to significant transcriptional variance with respect to cell-structure-related genes. Myotube formation requires the reorganization of subcellular architecture to implement the contractile properties necessary for mature muscle fiber function (Charge and Rudnicki, 2004). The three cytoskeletal-related gene sets tested in this analysis, 'cell structure,' 'cell motility' and 'cytoskeletal regulation by rho gtpase,' were each found to be significantly regulated based on their increased variance but were not identified as significantly enriched by either the KS or Wilcoxon tests. This result suggests that cytoskeletal modifications are accomplished not through the switching on or off of a single transcriptional program, but rather through a careful re-balancing of the structural protein complement expressed within the cells.

The Rho family of small GTPases is involved in a diverse array of structurally related cellular processes including regulation of actin cytoskeleton, cell polarity and microtubule dyamics (Bishop and Hall, 2000; Bokoch, 2000; Etienne-Manneville and Hall, 2002; Hall, 1998). These proteins are members of the Ras super family of small GTPases, and are similar to Ras proteins in size and sequence (Chardin, 1991). Expression and level of activation of distinct members of the Rho family starkly differs depending on the cell type and growth conditions, and recent evidence suggests the dynamic relationship between the Rho-GTPases during skeletal myogenesis (Takano *et al*., 1998), such that RhoA appears to be regulated by cell–cell adhesion and insulin/IGF signaling, both intrinsic to the myogenesis program. Other members of the Rho GTPase family, including Rac1 and Cdc42 are critical to myogenesis, although their specific role is unclear.

3.1.2 Ca^{++} signaling and transport Contraction of skeletal myofibrils is a fundamental mechanism that clearly defines this tissue from most others (Chin, 2005). Notably, the C2C12 myotubes are highly contractile, analogous to type 1 (fast twitch) myofibers, and therefore exhibit a gene expression signature that enables the integrity of this functional apparatus. Skeletal muscle contractility is stimulated by release of Ca^{++} from the sarcoplasmic reticulum into the cytosol. ATP-dependent pumps return Ca^{++} from the cytosol to the sarcoplasmic reticulum to lower cytosolic Ca^{++} levels and ultimately cease contraction. Given the central role of Ca^{++} signaling in muscle contractility, it is reasonable to find several Ca^{++}-related gene sets differentially regulated between contracting myotubes and non-contractile myoblasts. The available gene sets explicitly relating to calcium signaling and transport included 'calcium-mediated signaling' and 'calcium ion homeostasis,' each of which was detected by the LBF test (Supplementary Table 3). Genes related to Ca^{++} signaling in the context of functional skeletal muscle were up-regulated, such as troponin, adrenergic receptor beta 2, calsequestrin and triadin among others. However, genes related to Ca^{++} signaling in other cellular contexts such as a GPCR second messenger signaling were down-regulated, including calmodulin and various protein kinase C species. Such reciprocal regulation of Ca^{++} signaling mechanisms is critical in defining the intracellular state of the myocyte, and to maintain the appropriate and optimized cell phenotype.

3.1.3 Muscle-related signaling Several muscle-related signaling pathways were identified by the LBF analysis (Supplementary Table 3). The nicotinic acetylcholine receptor signaling gene set includes a subset of highly up-regulated genes directly involved in muscle contraction. Several cholinergic receptors that are found at the motor end plate are substantially up-regulated as are genes involved in the actin/myosin cross-bridge activities and a voltage-dependent Ca^{++} channel. In contrast, various myosin isoforms are down-regulated, which compromise the overall shift in pathway expression. The myosin complement of various muscle types including cardiac muscle, smooth muscle, as well as fast- versus slow-fiber-type skeletal muscle varies substantially. This myosin signature for a specific muscle type is important for specialized function, and hence the distribution within C2C12 myotubes involves up- and down-regulation of several myosin isoforms in order to achieve the correct balance that underlies contractile function.

The functional role of FGF2 signaling during the myogenic program is currently emerging in the literature. A role in wound healing has been demonstrated such that delivery of FGF2 to excisional muscle wounds enhanced skeletal muscle repair and triggered angiogenic responses that subsequently remodeled into arteriogenes (Doukas *et al*., 2002). At the myoblast level, FGF2 was shown to enhance proliferation of C2C12 cells and attenuate differentiation via activation of p44/p42-MAPK and suppression of Akt (Tortorella *et al*., 2001). Such data implicates FGF2 in the myogenic program and thus appearance of this pathway in the current data set is consistent, and notably this was identified by LBF analysis only.

3.2 Gene sets with low variance

One advantage the LBF test holds over other robust variance tests such as Fligner–Killeen (Conover *et al*., 1981) is that it maintains information about the relative magnitudes of the

variances tested. That is to say, LBF results describe not only if a gene set has different variance than the background, but also if its variance is larger or smaller than the background. Of the 48 gene sets returned by the LBF test, 7 have reduced variance. These gene sets (Supplementary Table 4) represent three categories of biochemical activities: protein biosynthesis, mRNA transcription and oxidative metabolism (Fig. 3).

Gene sets with tight variances do not easily lend themselves to interpretation. Although it is clear that these gene sets appear to be under tight transcriptional control, the purpose and mechanism of this control cannot be deduced from a single microarray experiment and warrants further investigation. Nevertheless, examination of the content of these gene sets offers one potential explanation. All three of these gene sets include large, multiunit complexes. These include *protein biosynthesis*: ribosomal proteins, mitochondrial ribosomal proteins, eukaryotic translation initiation factor, eukaryotic translation elongation factors; *oxidative metabolism*: ATP synthase, cytochrome *c* oxidase, NADH dehydrogenase, succinate dehydrogenase complex, ubiquinol cytochrome *c* reductase; *mRNA transcription*: CCR4-NOT transcription complex, polymerase (RNA) II (DNA directed). The stoichiometry of complex assembly may necessitate tighter transcriptional control in these pathways. For example, balanced transcription of ATP synthase subunits may be the most efficient way to maintain proper levels of the assembled protein complex. Unbalanced expression of a particular protein subunit might, in fact, have a deleterious effect on the overall pathway function via sequestration or other molecular dysfunction. In contrast, enzymatic or signaling pathways devoid of large complexes may make use of differentially regulating individual genes if their products participate in a rate-limiting step or serve as signal transducers or messengers.

These findings are in agreement with another recent study that identified the ribosome and oxidative phosphorylation gene sets among those with tightly correlated expression patterns across an independent data set (Huang *et al.*, 2006). Interestingly, the initial application of GSEA also identified oxidative phosphorylation as differentially regulated in muscle samples from type-2 diabetics (Mootha *et al.*, 2003). If the tight transcriptional patterns of these genes sets were to be observed across multiple experiments, one would need to consider the downstream effects on computational analyses. In principle, tightly coordinated gene sets would present cleaner signals more easily detected above background noise and would therefore be more amenable to detection by various enrichment-based analysis methods.

3.3 Gene sets identified by all three tests

A number of gene sets are identified as differentially regulated by all three tests (Supplementary Table 5). In general, these gene sets correspond to broad processes that are intimately involved in the differentiation process and undergo substantial modulation (see Fig. 4 for examples). These gene sets tend to be somewhat larger in size as well (means = 205.7, 61.2; $P = 0.018$). Given the size of these gene sets and the broad spectrum of processes they encompass, it is not surprising that these gene sets exhibit a number of different statistically significant

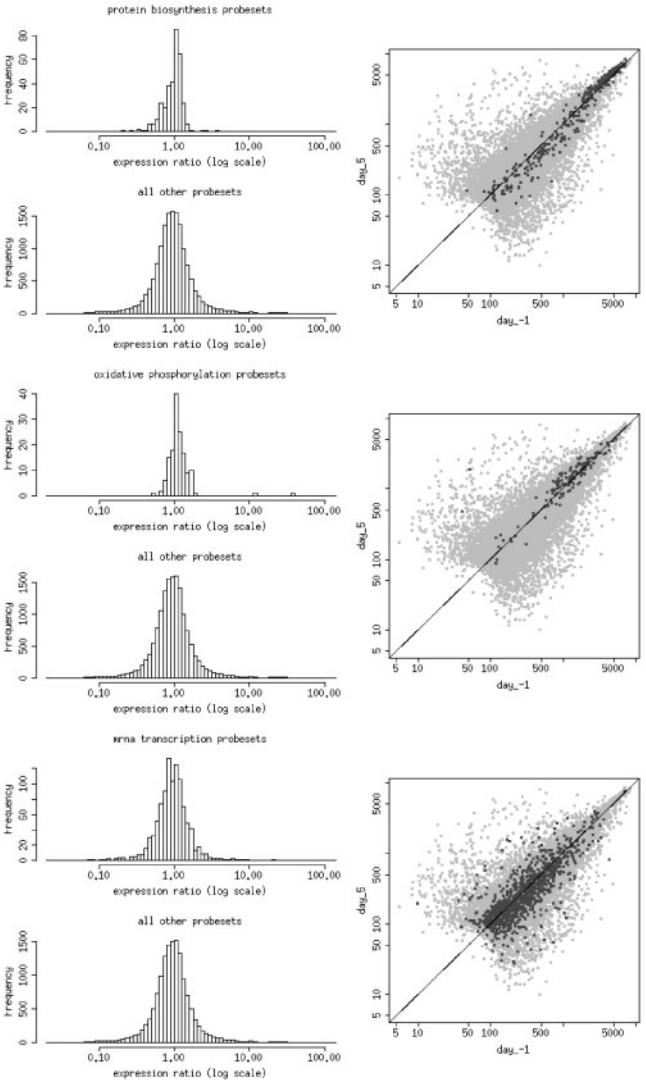

Fig. 3. The protein biosynthesis, oxidative phosphorylation and MRNA transcription gene sets all exhibit reduced transcriptional variance compared to background. Figure layouts are as in Figure 2.

rearrangements. It is likely that each of these sets contains positive and negative regulators of these processes and correspond to a number of sub-processes or pathways that function independently of each other. Several gene sets including muscle contraction (top panel) and cell communication (bottom panel) exhibit transcriptional changes that are large in magnitude and broad in scope. Such gene sets are detected as significantly modulated by all three statistical tests.

4 CONCLUSION

We have described an extension of the pathway-based enrichment framework to include a test for the variance of transcriptional responses. Application of the LBF test extends the repertoire of detectable transcriptional responses beyond

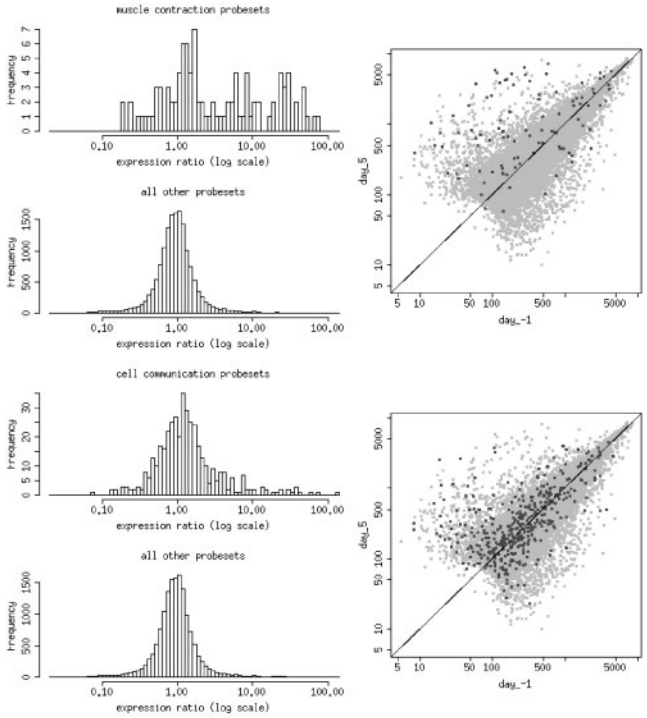

Fig. 4. Several gene sets including muscle contraction (top panel) and cell communication (bottom panel) exhibit transcriptional changes that are large in magnitude and broad in scope. Such gene sets are detected as significantly modulated by all three statistical tests. Figure layouts are as in Figure 2.

simple coordinated shifts to account for the complexities of biological processes as well as the vagaries of existing pathway knowledge. When applied to microarray data that captures the transcriptional shifts accompanying myogenesis, the LBF test unearths additional differentially regulated pathways previously overlooked by a standard enrichment analysis. Specifically, the LBF test implicated the modulation of processes related to Ca^{++} signaling, cell structure and muscle-specific signaling during C2C12 mouse skeletal myoblast differentiation. These responses tend to contain similar numbers of up- and down-regulated genes and were therefore missed by the standard enrichment analysis. The calcium-mediated signaling results illustrate the LBF test's ability to identify modulations in the presence of imperfect or loosely annotated gene sets. Although the calcium-mediated signaling gene set is accurately labeled, it includes genes involved in two distinct biological processes: muscle contractile signaling and second messenger signaling. Nevertheless, the LBF test picks out this gene set because of the reciprocal behavior of the genes that drive these two mechanisms. In contrast, both the standard enrichment analysis and KS test fail to detect these patterns. Instead these tests bump into a common limitation of available pathway databases which sometimes lump together related but separate processes into one category. This result suggests a method for improving pathway annotations. If a pathway such as 'calcium-mediated signaling' were to consistently show modulation across a number of different conditions via variance tests it may be worthwhile to dissect such a pathway into smaller, coordinately transcribed subsets. In this way, the analysis of several experiments would serve to feedback into the reservoir of available biological knowledge and facilitate more detailed and precise analyses of future experiments.

The application of a variance test is accompanied by a downstream need for more careful interpretation of results. The standard location-based enrichment technique identifies simple transcriptional patterns that lend themselves to easy interpretation: a particular pathway is either coordinately up- or down-regulated. In contrast, a significant LBF test *P*-value indicates that a pathway is modulated without an indication of the nature of the modulation. As an example, the 'calcium-mediated signaling' result clearly indicates that calcium homeostasis genes are being differentially modulated. It is only upon closer inspection of the constituent genes that we recognize this is a result of activation of muscle-related genes and deactivation of genes involved in other types of calcium signaling in this experiment. In another biological system, we might see a reciprocal behavior with activation of calcium second messenger signaling genes and inactivation of calcium muscle contraction genes. Both experiments would return a significant LBF result predicated on very different underlying biological phenomena.

The introduction of pathway-based analysis methods has provided a definitive step forward in microarray data mining. Through application of the LBF variance test, we have demonstrated the extensible nature of this framework. It should be noted that while the LBF test enhances the number and types of detectable biological responses, we believe there is still considerable room for the application of additional statistical tests in the use of pathway-based methods. Both tests used above identify fairly simple patterns. It is our belief that more complex patterns remain to be culled. As evidence, consider the results from the KS test which include nine gene sets whose modes of regulation are not explained by either the Wilcoxon or LBF tests. We foresee several natural avenues for the evolution of these pathway-based approaches. In the short term, we expect additional types of statistical tests may be designed to identify more complex or subtle patterns of regulation. These methods will advance in parallel with the increased knowledge base of biological pathways. While better delineation of pathways will improve these methods in the short term, one can hope that increased understanding of pathway topologies will lend itself to a broad application of more sophisticated computational techniques to high-throughput data sets.

ACKNOWLEDGEMENTS

The authors thank Penelope Kosinski for discussions motivating the application of a variance test; Leah Martell and Mathis Thoma for advice on statistical tests; Qicheng Ma for general methodological discussions; Stephen Elliman for feedback on various gene sets; the anonymous reviewers for their constructive feedback and suggestions.

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al*. (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, **21**, 2988–2993.

Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Bishop,A.L. and Hall,A. (2000) Rho GTPases and their effector proteins. *Biochem. J.*, **348 (Pt. 2)**, 241–255.

Bokoch,G.M. (2000) Regulation of cell function by Rho family GTPases. *Immunol. Res.*, **21**, 139–148.

Chardin,P. (1991) Small GTP-binding proteins of the ras family: a conserved functional mechanism? *Cancer Cells*, **3**, 117–126.

Charge,S.B. and Rudnicki,M.A. (2004) Cellular and molecular regulation of muscle regeneration. *Physiol. Rev.*, **84**, 209–238.

Chin,E.R. (2005) Role of Ca^{2+}/calmodulin-dependent kinases in skeletal muscle plasticity. *J. Appl. Physiol.*, **99**, 414–423.

Conover,W.J. *et al*. (1981) A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, **21**, 351–361.

Curtis,R.K. *et al*. (2005) Pathways to the analysis of microarray data. *Trends Biotechnol.*, **23**, 429–435.

Doniger,S.W. *et al*. (2003) MAPPFinder: using gene ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.*, **4**, R7.

Dopazo,J. (2006) Functional interpretation of microarray experiments. *Omics*, **10**, 398–410.

Doukas,J. *et al*. (2002) Delivery of FGF genes to wound repair cells enhances arteriogenesis and myogenesis in skeletal muscle. *Mol. Ther.*, **5 (5 Pt. 1)**, 517–527.

Draghici,S. *et al*. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Etienne-Manneville,S. and Hall,A. (2002) Rho GTPases in cell biology. *Nature*, **420**, 629–635.

Goeman,J.J. *et al*. (2005) Testing association of a pathway with survival using gene expression data. *Bioinformatics*, **21**, 1950–1957.

Hall,A. (1998) Rho GTPases and the actin cytoskeleton. *Science*, **279**, 509–514.

Huang,R. *et al*. (2006) Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen. *Genomics*, **87**, 315–328.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Khatri,P. *et al*. (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res.*, **32**, W449–W456.

Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.

Kong,S.W. *et al*. (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.

Mootha,V.K. *et al*. (2003) PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Siegal,S. (1956) *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Storey,J.D. and Tibshirani,R. (2003a) Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol. Biol.*, **224**, 149–157.

Storey,J.D. and Tibshirani,R. (2003b) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Szustakowski,J.D. *et al*. (2006) Identification of novel pathway regulation during myogenic differentiation. *Genomics*, **87**, 129–138.

Takano,H. *et al*. (1998) The Rho family G proteins play a critical role in muscle differentiation. *Mol. Cell. Biol.*, **18**, 1580–1589.

Tavazoie,S. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Tian,L. *et al*. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.

Tomfohr,J. *et al*. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.

Tortorella,L.L. *et al*. (2001) Critical proliferation-independent window for basic fibroblast growth factor repression of myogenesis via the p42/p44 MAPK signaling pathway. *J. Biol. Chem.*, **276**, 13709–13717.

Zahn,J.M. *et al*. (2006) Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Genet.*, **2**, e115.