Outlier Gene Set Analysis Combined with Top Scoring Pair Provides Robust Biomarkers of Pathway Activity

Michael F. Ochs[1,⋆], Jason E. Farrar[2], Michael Considine[1], Yingying Wei[3], Soheil Meschinchi[4], and Robert J. Arceci[5]

[1] The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, USA
mfo@jhu.edu
[2] College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, USA
[3] The Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
[4] Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[5] Ronald A. Matricaria Institute of Molecular Medicine, Phoenix Children's Hospital, Phoenix, AZ, USA

Abstract. Cancer is a disease driven by pathway activity, while useful biomarkers to predict outcome (prognostic markers) or determine treatment (treatment markers) rely on individual genes, proteins, or metabolites. We provide a novel approach that isolates pathways of interest by integrating outlier analysis and gene set analysis and couple it to the top-scoring pair algorithm to identify robust biomarkers. We demonstrate this methodology on pediatric acute myeloid leukemia (AML) data. We develop a biomarker in primary AML tumors, demonstrate robustness with an independent primary tumor data set, and show that the identified biomarkers also function well in relapsed AML tumors.

1 Introduction

The development of cancer is known to be driven by deregulation of several biological processes, referred to as the Hallmarks of Cancer [4], and loss of control of each process is required for the development of lethal cancers in almost all cases. Regulation of most of these Hallmarks relies on proper functioning of cell signaling pathways [5], which comprise sets of signaling proteins, primarily kinases and phosphatases, that work to transduce a signal through a cell by means of post-translational modifications of proteins. The deregulation of any single pathway can be driven by a mutation or other change in a single protein within the pathway [11].

⋆ Corresponding author.

A. Ngom et al. (Eds.): PRIB 2013, LNBI 7986, pp. 47–58, 2013.

1.1 Outlier Gene Set Analysis

The dominance of pathways over genes in the etiology of cancer creates a problem for statistical analysis that focuses on determining global behaviors in cancers in general or types of cancer in particular. Since loss of regulation of a pathway is the critical event, but global measurements focus on genes and the proteins they encode, there is a mismatch in the statistic (based on data from genes) and the effect (based on pathway deregulation). This suggests a need for a pathway-based statistic for use in cancer studies.

The first issue to resolve is that any given gene in a subtype of cancer is likely to be affected in only a small fraction of individuals, since there are many potential genes that may drive pathway deregulation. For example, the well-studied RAS-RAF pathway may become deregulated through overexpression of the EGFR receptor, mutation of the RAS, RAF, or MAPK genes, or mutation or overexpression of the MYC transcriptional regulator. Any individual is likely to have only one such change, and no single change is likely to rise above $\sim 50\%$ of cases, with most lying between 5% and 15%. This limits the value of standard statistical tests, such as t-tests or ANOVA analyses.

However, outlier analysis, such as Cancer Outlier Profile Analysis [9], provides a method to identify those genes that are deregulated in only a subset of individuals. While useful, this alone will not provide the required identification of deregulated pathways, although it should provide an indication of significance of the individual pathway members. With Gene Set Analysis (GSA) we can integrate these estimates of significance to provide an overall estimate of pathway significance on a global scale, which we refer to as Outlier Gene Set Analysis (OGSA). This provides a global estimate of pathway deregulation in cancer subtypes.

1.2 Pathway-Based Top Scoring Pairs

The fundamental measurements we make clinically remain linked to genes, not pathways. This complicates the development of diagnostic tests for the drivers in cancer, the pathways. In general, we visualize the deregulation of the pathway through heatmaps and other data-driven visualization tools. However, these provide poor clinical utility as the results change with addition of data, making them inappropriate for clinical tests that must deduce a probability from an isolated measurement, and they have been shown to be strongly platform dependent, increasing the potential cost and reducing the opportunity for innovation.

In order to create a method that could identify robust potential biomarkers, the multigene signature generated from discriminant analysis can be replaced by pairs of genes that change their relative level of expression [14], known as a Top Scoring Pair (TSP) [1]. In TSP, the statistic of interest is how well the measurements on a pair of genes distinguish two classes, relying on the inversion of the values of measurements between classes. This provides a normalization-independent approach that makes switching measurement technologies far more likely to succeed [12]. However, a limitation of TSP is that it searches through all

possible TSPs, introducing the potential of chance identification of pairs that are not robust and fail to validate. Instead, we build a pathway TSP set by limiting the domain for generating TSPs to pathways of interest in the set of statistically significant pathways generated by OGSA. In this way, we focus the methodology on biologically-motivated gene sets, more suitable for clinical development than unbiased discovery.

1.3 Pediatric AML and the TARGET Initiative

Acute Myeloid Leukemia (AML) is a cancer of the blood affecting roughly $15,000$ individuals per year in the USA, and childhood patients show $\sim 60\%$ five year survival. However, the outcomes are highly dependent on karyotype-defined subtype, and initiatives to improve care for pediatric patients have led to broad molecular studies through the NCI Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative.

1.4 Outline of Paper

In this paper, we describe the methodology in sections 2.1 and 2.2 together with the analysis of the AML data in section 2.3. In section 3, we show that OGSA of TARGET promoter methylation data identified the Hedgehog signaling pathway and the Cytochrome P450 metabolic pathway as highly epigenetically deregulated in pediatric AML. Using only genes associated with these pathways for the development of a set of TSPs, we demonstrate that we obtained a robust signature of pathway deregulation that was significant in an independent data set and also significant in samples from individuals whose cancer relapsed. Importantly, this suggests a novel therapeutic strategy in these patients and provides a potential treatment biomarker for this therapy.

2 Methods

Overall we adopted a number of key methodologies developed for identifying outlier genes and generating robust TSPs. We integrated these methods into a pathway-centric statistical approach that leverages outlier statistics to generate pathway statistics through OGSA and generates TSPs related to key pathways.

2.1 Outlier Gene Set Analysis

The standard method employed in cancer research for outlier analysis is Cancer Outlier Profile Analysis [9], which generates statistics by comparing the outlier distributions to an empirical null generated by permutation of class labels. However, this is computationally expensive and, importantly, we required only the rank of the genes and not their significance, since we utilized a rank-based gene set test (see below). Thus, we generated statistics using a modification that permits rapid p-value estimation, although this estimation is in general less reliable than that generated by a permutation test.

Each observation in a case sample was compared to the empirical distribution of expression values of the same gene for control samples following a ranksum methodology [3]. For gene g, we calculated the right-tail empricial p-value as

$$\hat{p}_{gt} = \frac{1}{N_{p0}} \sum_{i=1}^{N_{p0}} I(X_{gt} \leq X_{gi}) \tag{1}$$

where we indexed the control samples with i and the case samples by t with N_{p0} control samples and N_{p1} case samples. The corresponding left-tail empirical p-value was calculated as

$$\hat{p}_{gt} = \frac{1}{N_{p0}} \sum_{i=1}^{N_{p0}} I(X_{gt} \geq X_{gi}). \tag{2}$$

For both cases, we generated a $G \times N_{p1}$ matrix of empirical $\hat{p}$-values for each gene as an outlier in each case sample.

We modified these equations slightly in this study to incorporate biological knowledge of the impact of changes in methylation. Because cancers often show global methylation changes involving loss of intergenic methylation and increased methylation near genes, including areas measured by array technologies, it is not unusual for almost all tumor samples to show a slight increase in methylation in gene promoters relative to normal samples. However, these small methylation changes are not meaningful biologically, as they are not enough to drive changes in expression of the genes. As such, we modified Equations 1 and 2 by replacing X_{gi} by $X_{gi} + 0.1$ and $X_{gi} - 0.1$ respectively. Effectively, we counted outliers only when there was at least a 10% change in the level of methylation.

To generate rank statistics, we converted the $\hat{p}$-values to an indicator of significance by testing them against a Bonferroni corrected $\alpha = 0.05$ by

$$\hat{m}_{gt} = I(\hat{p}_{gt} \leq \frac{\alpha}{N_{p1}}) \tag{3}$$

where 1 indicates significant at level α and 0 indicates insignificant. The rank statistic was the sum of the indicator across all case samples, effectively ranking genes from N_{p1} to 0.

We analyzed these rank statistics using a mean rank gene set enrichment test [10], as provided in the limma R package [13], comparing the statistics of the genes in a gene set to genes outside the set. The mean-ranks of the test statistics for the genes were used for comparison, which matched our use of only the ranks of genes from outlier analysis. Gene sets were defined by the KEGG and BioCarta pathways [6] and final p-value estimates on the pathways were corrected for multiple testing using the Bonferroni method.

2.2 Pathway-Based Top Scoring Pairs

The OGSA method provides pathways that are significantly different between cases and controls, but it does not provide a suitable methodology for the

development of a test for a new sample. In order to generate such a test, we applied OGSA to highlight pathways of interest. We refined significant pathways by inspection, focusing on suitability for drug targeting or removal of pathways either universally modified or already addressed in treatment. We then used only the genes associated with the refined pathway list in TSP (i.e., those genes that define the gene set for this pathway in KEGG).

Table 1. Example TSP

	$G_i < G_j$	$G_i > G_j$	
Case	N_{TP}	N_{FN}	N_{case}
Control	N_{FP}	N_{TN}	N_{control}
	N_{callCase}	$N_{\text{callControl}}$	N

The choice of a TSP reduces to maximization of prediction in a Fisher two-way table, such that Table 1 provides the best possible predictive value for the measured levels G, here promoter methylation, of two genes i and j, where the relative levels of these genes determines the result of the test, with $G_i < G_j$ predicting a case and the inverse a control. The TSP is determined by finding the pair of genes that maximizes

$$\Delta_{ij} = |\frac{N_{TP}}{N_{\text{case}}} - \frac{N_{FP}}{N_{\text{control}}}|, \tag{4}$$

where N_{TP} is the number of true positives, N_{FN} is the number of false negatives, N_{FP} is the number of false positives, and N_{TN} is the number of true negatives. The total number of measurements is N, divided into N_{case} cases and N_{control} controls. As TSP does not always provide ideal separation due to the inherent complexity of the underlying biology, the extension to kTSP, where multiple TSPs vote on case or control status, is natural [14].

Here we used kTSP, as implemented in the R ktspair package [2]. We generated five TSPs in our training set for voting on status of the samples.

2.3 Analysis of TARGET Methylation Data

We applied the OGSA and TSP methods to a set of samples from the NCI TARGET initiative. The data comprised 192 diagnostic samples of pediatric AML, 192 remission samples from the same patients after frontline treatment, and 46 relapse samples from those patients with a recurrence of AML. All measurements were made with Illumina HumanMethylation27 BeadChip arrays, and beta values (percent methylation) were generated from U and M probes. Methylation estimates showing low variance across all samples were removed, leaving 19999 promoter methylation estimates associated with 11871 genes. A training set of diagnostic and remission samples was generated from 96 patients by choosing

roughly 50% of the samples of each karyotype in the data set. This balanced set was chosen to avoid biasing the training set to any particular diagnostic subtype, as different karyotypes have different outcomes in AML. Samples from the remaining 96 patients formed the test set, and an additional set of remission and relapse samples was generated based on the 46 relapse samples.

The OGSA method was applied to the training set and significant pathways were determined. For genes with multiple associated methylation probes, the probe with the highest mean methylation was retained. From the significant pathways, one driven by hypermethylation and one by hypomethylation were chosen based on the usefulness for drug targeting and metabolism as well as on their lack of being associated globally with all cancers. The use of one hypermethylation and one hypomethylation driven pathway was to increase the potential range of top-scoring pairs.

Five TSPs were generated from the probes associated with the genes assigned to the two pathways using the ktspair package applied to the training data set. These pairs were then used to vote on each sample, and the cutoff that maximized the predictive power of the pairs was used. These same pairs and cutoff were then applied to the training data set and to the relapse-remission data.

The key targetable pathway was also visualized using a heatmap of the genes in the pathway. This permitted visual comparison of the separation of diagnostic samples from remission samples, as well as the separation of relapse and remission samples. To test whether the pathway associated with karyotype, separation of karyotype on the heatmap was also investigated; however, there was no correlation (heatmap not shown).

3 Results

We applied our methods to the TARGET AML data comprising 430 samples as discussed in the Methods section. We analyzed the three separate data sets, Training, Test, Relapse, as follows. We first performed outlier analysis on the Training data, ranking all genes based on their outliers according to the sum across all diagnosis samples (Equation 3). These gene ranks were used to generate a set of significant pathways from the KEGG and Biocarta pathways using OGSA. We focused on two pathways from this set, the KEGG Hedgehog Signaling and Cytochrome P450 Metabolism pathways, for reasons detailed below. Using genes from the Hedgehog pathway, we created heatmaps of the Training, Test, and Relapse data to visualize the separation of samples. Using only the Training data, we then created five TSPs from these pathways. We tested these TSPs on the Test and Relapse data, using an assumption that a vote for a diagnostic sample was equivalent to a vote for a relapse sample in the test.

3.1 Outlier Analysis and Gene Ranks

Outlier analysis according to Equation 3 provided outlier ranks for all genes. As shown in Figure 1, highly ranked genes showed substantial increases in

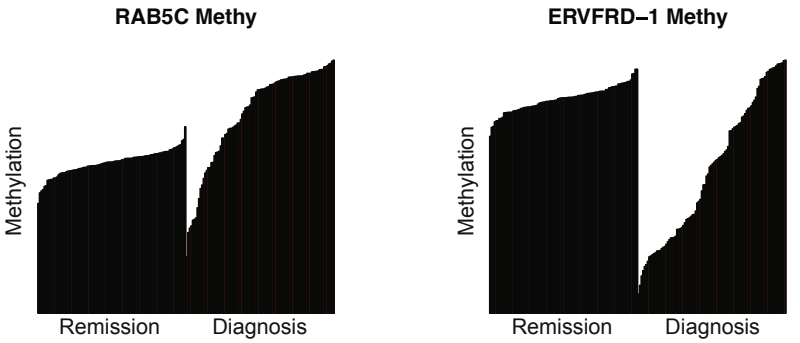

Fig. 1. The highest ranking right-tail and left-tail outlier genes from the Training data

methylation in diagnostic samples relative to remission samples. The top-ranked right-tail outlier gene, RAB5C, had 53 of 92 diagnostic samples called outliers, while the left-tail outlier gene, ERVFRD, had 48 of 92 diagnostic samples called outliers. The gene-based statistic is then provided by the rank from the gene with most outliers to the one with the fewest. The right-tail and left-tail rank lists were used in OGSA separately.

3.2 Significant Pathways from OGSA

The results of OGSA analysis of the KEGG and Biocarta pathway genes sets from the MSigDB database [8] are presented in Table 2. The p-values are Bonferroni corrected values from the mean rank gene set test. All pathways with significant p-values at the traditional $\alpha = 0.05$ are included in the table.

Many pathways in the right-tail analysis are seen in most GSA analyses of cancer data, including those involving focal adhesion and extracellular matrix receptor signaling (KEGG ECM Receptor Interaction, Cell Adhesion Molecules, Focal Adhesion pathways), pathways related to cancer (KEGG Neuroactive Ligand Receptor Interaction, Basal Cell Carcinoma, Pathways in Cancer pathways), and sets that appear significant in cancer studies due to the presence of genes related to integrin signaling and MAPK pathway activity (KEGG Dilated Cardiomyopathy and Arrhythmic Right Ventricular Cardiomyopathy pathways). These processes are deregulated in most cancers and do not provide novel insights to AML.

The pathways in the left-tail analysis are primarily involved in metabolism or immune responses. These pathways, in general, do not provide useful information for treatment and are generally hard to interpret in terms of cancer biology. Note that KEGG Neuroactive Ligand Receptor Interaction is significant in the left-tail analysis and the right-tail analysis, which indicates that methylation

Table 2. Significant KEGG and Biocarta Pathways

Right-Tail Outlier Results	p-Value
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	< 0.00001
KEGG ECM RECEPTOR INTERACTION	< 0.00001
KEGG HEDGEHOG SIGNALING PATHWAY	0.00001
KEGG BASAL CELL CARCINOMA	0.00032
KEGG PATHWAYS IN CANCER	0.00061
KEGG CELL ADHESION MOLECULES CAMS	0.00226
KEGG DILATED CARDIOMYOPATHY	0.00277
KEGG CALCIUM SIGNALING PATHWAY	0.01343
KEGG FOCAL ADHESION	0.01562
KEGG ARRHYTH RT VENTR CARDIOMYOPATHY ARVC	0.03704
Left-Tail Outlier Results	p-Value
KEGG COMPLEMENT AND COAGULATION CASCADES	< 0.00001
BIOCARTA COMP PATHWAY	< 0.00001
KEGG OLFACTORY TRANSDUCTION	< 0.00001
KEGG DRUG METABOLISM CYTOCHROME P450	0.00001
KEGG LINOLEIC ACID METABOLISM	0.00001
BIOCARTA CLASSIC PATHWAY	0.00004
KEGG METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.00010
KEGG TYROSINE METABOLISM	0.00014
KEGG ARACHIDONIC ACID METABOLISM	0.00033
KEGG ETHER LIPID METABOLISM	0.00038
BIOCARTA LECTIN PATHWAY	0.00074
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	0.00301
KEGG STEROID HORMONE BIOSYNTHESIS	0.00348
KEGG RETINOL METABOLISM	0.01062
BIOCARTA INTRINSIC PATHWAY	0.03099

changes in the promoters of genes in this pathway include both hyper- and hypo-methylation.

The KEGG Hedgehog Signaling Pathway in the right-tail analysis attracted our attention, because Hedgehog signaling is known to be a driver of proliferation and antiapoptotic behavior, is involved in multiple cancers, is not typically associated with AML, and provides a potential target for treatment. To visualize the Hedgehog pathway methylation, we generated heatmaps of the samples, looking for separation of diagnostic, remission, and relapse samples (see Figure 2).

The Drug Metabolism Cytochrome P450 pathway and related Metabolism of Xenobiotics by Cytochrome P450 pathway in the left-tail analysis was suggestive given the importance of Cytochrome P450 in processing of therapeutic agents. The genes in this pathway coupled to the Hedgehog pathway genes provided a set of hypermethylated and hypomethylated genes suitable for creating a biomarker using TSP.

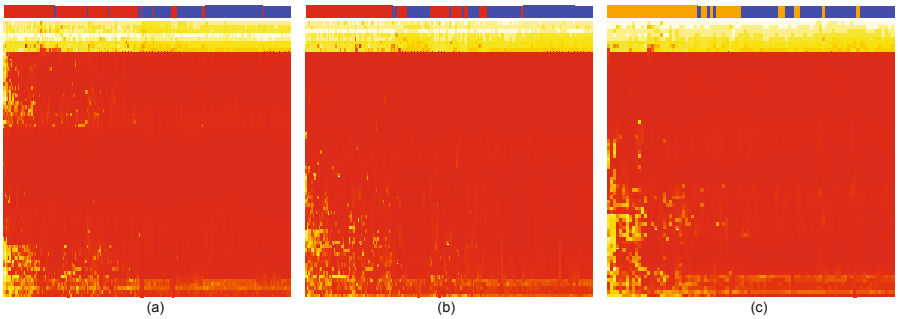

Fig. 2. Heatmaps of the methylation levels for promoters of genes in the KEGG Hedgehog pathway across patients in (a) the Training data, (b) the Test data, and (c) the Relapse data. In the top bar, blue indicates a remission sample, red a diagnostic sample, and orange a relapse sample. Genes are in rows and patients in columns. Yellow indicates high methylation ($\beta \to 1$) and red low methylation ($\beta \to 0$).

3.3 kTSP Classifiers for Hedgehog and Cytochrome P450 Pathways

In order to create a robust methylation signature for the Hedgehog and Cytochrome P450 pathways, we applied the kTSP algorithm to a subset of the Training data limited to promoter methylation levels of genes in the Hedgehog Signaling and Cytochrome P450 Metabolism pathways. We identified a set of 5 pairs that discriminate the diagnostic samples from the remission bone marrow samples (see Figure 3 where colors match the upper bar in Figure 2, so that blue is a remission sample and red a diagnostic sample). As seen in Table 3, this provided excellent prediction on the training set, with $p < 2.2 \times 10^{-16}$ and an odds ratio of 81 with a 95% confidence interval of $[28, 294]$.

Applying this signature to the Test data resulted in excellent prediction of diagnostic vs. remission samples, with $p < 2.2 \times 10^{-16}$, and an odds ratio of 128 with a 95% confidence interval of $[40, 563]$. Interestingly, the application of the same signature to the Relapse data set was also predictive, now of relapse vs. remission, with $p = 1.8 \times 10^{-6}$, and an odds ratio of 15 with a 95% confidence interval of $[4, 87]$. This suggests that relapse in pediatric AML may be partially driven by recurrence of methylation changes in the promoters of Hedgehog Signaling and Cytochrome P450 metabolism pathway genes, although the drop in sensitivity suggests that the relapse samples may be more diverse in this methylation than the diagnosis samples. Importantly, all tests show excellent Positive Predictive Values (94%, 95%, and 89% respectively), as is desirable for a test that could define treatment, since the vast majority of positive tests are related to positive pathway status.

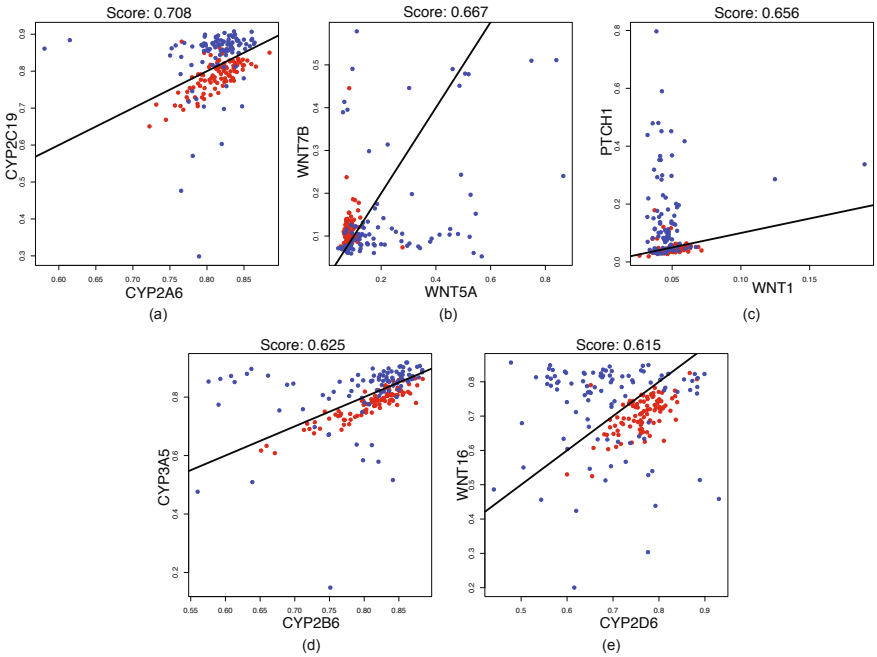

Fig. 3. The Five Top Scoring Pairs used to generate Table 3

Table 3. kTSP Classifier Performance

Training	Dx	Rm	**Test**	Dx	Rm	**Relapse**	Rl	Rm
Call Dx	79	5		82	4	Call Rl	24	3
Call Rm	17	91		14	92	Call Rm	22	43

4 Discussion

The coupling of outlier statistics, gene set analysis, and top scoring pair methods provides a solid methodology to identify deregulated pathways in cancer and to define a robust signature of their activity. We have shown that the method determines a robust marker, here comprising five TSPs, that validates in a completely novel data set, albeit one measured on the same platform at the same institution. Intriguingly, the marker does predict activity in the pathway in a subset of the relapse samples, suggesting both robustness of the marker and, potentially, that relapsed pediatric AML shows more heterogeneity than primary pediatric AML in Hedgehog activity. However, this suggestion is tempered by the low numbers and the known mismatch in karyotypes between primary and recurrent AML,

even though there was no correlation of Hedgehog pathway methylation with karyotype in primary tumors.

AML, specifically, and cancer in general, is difficult to treat effectively in most cases. Natural heterogeneity in response to treatment likely arises from both differences in molecular tumor characteristics and differences in systemic responses of individual patients [7]. Given this complexity, methods to define robust markers of potentially targetable pathways are extremely valuable to guiding treatment decisions, since the absence of cancer driver pathway activity should contraindicate targeted treatments for that pathway. The Positive Predictive Values (PPVs) from this test are therefore particularly promising, since a positive test is strongly indicative of pathway activity.

There remains a great need for more powerful, guided computational methods in cancer research and treatment. The complexity of the biological systems and a massive curse-of-dimensionality issue driven by small sample size coupled to genome-wide measurements of multiple molecular species present a formidable challenge requiring nonlinear modeling and novel computational learning techniques. It is likely the only viable approach will be to accept higher bias to reduce variance, and we have presented one such approach, where we limit our biomarker search based on statistically significant but knowledge-refined pathways.

Acknowledgements. MFO was funded by NIH/NLM R01LM011000 and NIH/NCI CCSG P30CA006973. MFO, JEF, MC, SM, and RJA were funded by the NIH/NCI U01 CA097452 National Childhood Cancer Foundation (TARGET). YW was partially funded by NIDCR RC1DE020324. RJA also received support from the endowed King Fahd Chair in Pediatric Oncology.

References

1. Geman, D., d'Avignon, C., Naiman, D.Q., Winslow, R.L.: Classifying gene expression profiles from pairwise mrna comparison. Statistical Applications in Genetics and Molecular Biology 3(1), 19 (2004)
2. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10), R80 (2004)
3. Ghosh, D.: Discrete nonparametric algorithms for outlier detection with genomic data. Journal of Biopharmaceutical Statistics 20(2), 193–208 (2010)
4. Hanahan, D., Weinberg, R.A.: The hallmarks of cancer. Cell 100(1), 57–70 (2000)
5. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. Cell 144(5), 646–674 (2011)
6. Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A.: The KEGG databases at genomenet. Nucleic Acids Res. 30(1), 42–46 (2002)
7. Knox, S.S., Ochs, M.F.: Implications of systemic dysfunction for the etiology of malignancy. Gene. Regul. Syst. Bio. 7, 11–22 (2013)

8. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., Mesirov, J.P.: Molecular signatures database (MSigDB) 3.0. Bioinformatics 27(12), 1739–1740 (2011)
9. MacDonald, J.W., Ghosh, D.: COPA–cancer outlier profile analysis. Bioinformatics 22(23), 2950–2951 (2006)
10. Michaud, J., Simpson, K.M., Escher, R., Buchet-Poyau, K., Beissbarth, T., Carmichael, C., Ritchie, M.E., Schütz, F., Cannon, P., Liu, M., Shen, X., Ito, Y., Raskind, W.H., Horwitz, M.S., Osato, M., Turner, D.R., Speed, T.P., Kavallaris, M., Smyth, G.K., Scott, H.S.: Integrative analysis of RUNX1 downstream pathways and target genes. BMC Genomics 9, 363 (2008)
11. Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.M., Gallia, G.L., Olivi, A., McLendon, R., Rasheed, B.A., Keir, S., Nikolskaya, T., Nikolsky, Y., Busam, D.A., Tekleab, H., Diaz Jr., L.A., Hartigan, J., Smith, D.R., Strausberg, R.L., Marie, S.K., Shinjo, S.M., Yan, H., Riggins, G.J., Bigner, D.D., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V.E., Kinzler, K.W.: An integrated genomic analysis of human glioblastoma multiforme. Science 321(5897), 1807–1812 (2008)
12. Price, N.D., Trent, J., El-Naggar, A.K., Cogdell, D., Taylor, E., Hunt, K.K., Pollock, R.E., Hood, L., Shmulevich, I., Zhang, W.: Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. Proc. Natl. Acad. Sci. U S A 104(9), 3414–3419 (2007)
13. Smyth, G.K.: Limma: linear models for microarray data, pp. 397–420. Springer, New York (2005)
14. Tan, A.C., Naiman, D.Q., Xu, L., Winslow, R.L., Geman, D.: Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 21(20), 3896–3904 (2005)