HHS Public Access

Author manuscript

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2015 May 22.

Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2014 ; 11(3): 520–532. doi:10.1109/TCBB.2013.153.

Outlier Analysis and Top Scoring Pair for Integrated Data Analysis and Biomarker Discovery

Michael F. Ochs [Senior Member, IEEE],

Department of Mathematics and Statistics, The College of New Jersey, Ewing, NJ, USA

Jason E. Farrar,

College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, USA

Michael Considine,

School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Yingying Wei,

Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

Soheil Meshinchi, and

Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Robert J. Arceci

Ronald A. Matricaria Institute of Molecular Medicine, Phoenix Children's Hospital, Phoenix, AZ, USA

Michael F. Ochs: ochsm@tcnj.edu

Abstract

Pathway deregulation has been identified as a key driver of carcinogenesis, with proteins in signaling pathways serving as primary targets for drug development. Deregulation can be driven by a number of molecular events, including gene mutation, epigenetic changes in gene promoters, overexpression, and gene amplifications or deletions. We demonstrate a novel approach that identifies pathways of interest by integrating outlier analysis within and across molecular data types with gene set analysis. We use the results to seed the top-scoring pair algorithm to identify robust biomarkers associated with pathway deregulation. 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 pediatric AML tumors.

Keywords

Genomics; Data integration; Statistical analysis; Biomarker

Correspondence to: Michael F. Ochs, ochsm@tcnj.edu.

Software

R code for all analysis and visualization tools presented here will be available through Bioconductor when the package is approved. Software is also available by request. Please email ochsm@tcnj.edu with OGSA in the subject line to request code.

Introduction

The development of cancer is known to be driven by deregulation of several biological processes, referred to as the Hallmarks of Cancer [1], 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 [2], which comprise sets of signaling proteins, primarily kinases and phosphatases, that work to transduce a signal through a cell by means of posttranslational 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 [3].

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 ~ 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.

Outlier analysis, such as Cancer Outlier Profile Analysis [4], 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.

Molecular Events Affecting Gene Function

The change in gene function that leads to pathway deregulation may arise in multiple different ways. The most studied change is a mutation that leads to either loss or gain of function, causing a pathway to activate in the absence of signal. However, overexpression of pathway members, amplification or loss of gene loci, and epigenetic silencing through methylation of promoters or the loss of such silencing have all been shown to play roles in pathway deregulation [5, 6, 7].

For each type of molecular event, a different genome-wide measurement methodology is applied, providing a global view of the data across tumors and normal samples. Integration

of these data is problematic, as each type has its own distribution and errors. We show, however, that outlier analysis provides a reliable integration approach, since outliers for each data type can be identified within the measurements associated with that type and then integrated across the different types by simply summing counts, so long as the desired result is a ranking of genes, as it is here, and not an estimate of significance. Significance can be estimated by generation of an empirical p -value through permutation tests [8], if desired, although with high computational costs for large data sets.

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. Unfortunately, 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 of change platforms 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 [9], known as a Top Scoring Pair (TSP) [10]. 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 [11]. However, a limitation of TSP is that it searches through all possible pairs of genes, introducing the potential of chance identification of pairs that are not robust and fail to validate. Here 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 for unbiased discovery.

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 ~60% five year survival. However, the outcomes are highly dependent on karyotype-defined subtype [12], 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.

Pathway Analysis

There have been a number of methodologies developed for identifying deregulated pathways, particularly in cancer. These methods primarily rely on interactome and gene expression data. K-Q Liu *et al* built networks based on curated pathway information and interactome data and then identified the most deregulated pathways from standardized expression data [13]. From the pathways that best distinguished cancer and normal samples,

they built biomarkers. Y Liu et al introduced GIENA, which builds a gene-gene metric based on a four dimensional vector describing the interaction of each pair of genes. Deregulated pathways are then identified from curated pathway sets using a Z-statistic based on the gene vectors with significance determined by permutation testing [14]. Similarly, our own previous work applied Bayesian Decomposition, a non-negative matrix factorization algorithm, to generate Z-scores for transcriptional regulators from expression data and demonstrated an ability to detect pathway activity changes in response to targeted therapy [15]. Kim *et al* used eQTL analysis on molecular interaction data to identify changes in networks [16].

Ulitsky *et al* introduced DEGAS, which coupled case-control expression analysis to network analysis to identify deregulated subnetworks [17]. This method is similar to what we present here in that it aims to create biomarkers for cancer related to specific signaling aberrations. Our work relies on standard outlier tests for identifying the statistically significant outliers, limits the tests only to curated pathways not subnetworks, and relies on TSP in order to generate robust biomarkers. However, the spirit of the two methods is quite similar.

Outline of Paper

In this paper, we describe the methodology in sections 2.1 through 2.3 together with the analysis of the AML data in sections 2.4 through 2.6. In section 3, we show that OGSA of TARGET promoter methylation data identified the Hedgehog signaling pathway as highly epigenetically deregulated in pediatric AML. Using only genes associated with this pathway 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. In section 4, we show how to integrate data from different molecular domains using OGSA on copy number, methylation, and expression data simultaneously. We also provide visualization of tumor-specific outliers, demonstrating how the method can potentially provide personalized identification of targets for treatment.

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. A flowchart for the method is shown in Fig. 1 and the specifics for each step are outlined below, beginning with the methodologies shown in rounded rectangles in the figure.

Outlier Statistics and Counts

The standard method employed in cancer research for outlier analysis is Cancer Outlier Profile Analysis [4], which generates statistics by comparing the outlier distributions to an empirical null generated by permutation of class labels. The method has been modified

slightly by Tibshirani and Hastie [18], and we encoded their method as an option within our OGSA method.

However, both these methods have limitations when counting outliers. It is often the case that the distribution of medians and median absolute deviations (MADs) permits outliers to be called in cases where the deviations are insignificant biologically. As such, we have also implemented a rank sum outlier approach, modified from Ghosh [19], where we set minimum change levels for the calling of an outlier. This eliminated many outliers where the change was not biologically meaningful (e.g., methylation percent change of less than two percent). Such a change would be difficult to implement within the other outlier methods.

For the Tibshirani and Hastie outlier method, the measured values X for each gene were transformed to rescale all genes to the same distribution by

$$\tilde{X}_{ij} = \frac{X_{ij} - \text{median}(X_i)}{\text{MAD}(X_i^{n \in N})} \quad (1)$$

for all genes i in all samples j , where the MAD is calculated only over the controls $n \in N$ (i.e., normal samples). These values were then used to determine outliers in the cases $t \in T$ as

$$\hat{m}_{it} = \tilde{X}_{it} > (\tilde{X}_i^{0.75} + IQR(\tilde{X}_i)) \quad (2)$$

$$\hat{m}_{it} = \tilde{X}_{it} < (\tilde{X}_i^{0.25} - IQR(\tilde{X}_i)) \quad (3)$$

where m is the indicator for an outlier in the right-tail (top) or left-tail (bottom). The superscripts give the quartile and IQR represents the interquartile range.

However, due to rescaling, there is no easy way to handle cases where the outliers are mathematically significant but arise from too small a difference to have an effect biologically. To resolve this issue, we used the empirical distribution of measured values of the same gene for control samples following a ranksum methodology. For gene i , we calculated the right-tail and left-tail empirical p -value as

$$\hat{p}_{it} = \frac{1}{N} \sum_{n=1}^N I(X_{it} \leq (X_{in} + X_0)) \quad (4)$$

$$\hat{p}_{it} = \frac{1}{N} \sum_{n=1}^N I(X_{it} \geq (X_{in} - X_0)) \quad (5)$$

respectively, where X_0 provides a biologically motivated minimum significant difference and $I(\cdot)$ is the indicator function that equals 1 when the comparison is true and 0 when false. We indexed the control samples with n and the case samples by t with N control samples and T case samples. For both cases, we generated a $G \times T$ matrix of empirical p -values for gene $i \in G$ as an outlier in case sample $t \in T$. Note that here an empirical p -value can be zero, so

this method should not be assumed to give meaningful p -values for uses beyond counting outliers.

For this study we set a data-type specific minimum difference of 1.0 for log2 expression, 0.1 for fraction methylation (i.e., γ values), and 0.5 for copy number variation. We investigated values for these of 0 – 4, 0 – 0.75, and 0 – 2.0, and we found only the extremes created significant differences in estimation of signaling activity changes.

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

$$\hat{m}_{it} = I(\hat{p}_{it} \leq \alpha) \quad (6)$$

where 1 indicates significant at level α and 0 indicates insignificant.

Outlier Gene Set Analysis and Tumor Specific Outlier Visualization

The $\hat{m}_{it}$ values are indicators of specific outliers. To generate ranks on genes for gene set analysis, we then performed summation across all cases for each gene as

$$R_i = \sum_{d=1}^D \sum_{t=1}^T \hat{m}_{it}^d \quad (7)$$

where R_i gives the rank for gene i and the first sum can be over an individual molecular data type $d \in D$ or over all D types (e.g., expression, methylation, and copy number). The rank statistic was the sum of the indicator across all case samples, effectively ranking genes from T to 0 for a single data type and DT to 0 for D data types.

The ranking of the genes in this manner assumes that there are an equal number of measurements on each gene, so that there is no bias introduced by missing data, where an outlier has no possibility of being detected in a sample for a gene. As such, data preprocessing requires the removal of either samples or genes to make sure that for each gene the same number of tumor samples are present across all platforms. To remove more subtle potential bias, we required that the genes are also measured in the same set of normal samples.

In contrast, for the case where not all samples were measured for all molecular types, it is possible to retain all measurements. The number of maximum calls is then

$$R_i = \sum_{d=1}^D \sum_{t=1}^{T_d} \hat{m}_{it}^d \quad (8)$$

where T_d now indicates the number of tumor samples for data type d . While this retains the validity of ranking of the genes, it can introduce bias towards a specific molecular type. A decision on whether to forego some samples in favor of less bias is dependent on the specific study. In this work, we had complete measurements in all diagnostic samples for each gene, so our integration automatically satisfied Eqn. 7.

We analyzed these rank statistics for promoter methylation and for the integrated data using a Wilcoxon gene set test as provided in the limma R package [20], comparing the rank of the genes in a gene set to genes outside the set. Gene sets were defined by the KEGG and BioCarta pathways [21] and final p -value estimates on the pathways were corrected for multiple testing using the Bonferroni method.

One advantage to the identification of tumor-specific outlier calls, $\hat{m}_{it}^d$, is that the potential causative molecular changes can be identified for a deregulated pathway within an individual tumor. Since $\hat{m}_{it}^d$ provides all outlier calls in all data types, the specific aberrations present in each pathway can be tabulated and visualized. We utilized the heatmap.2 function in the R package gplots to visualize a binary encoding of the outliers. For each gene in each tumor, we encoded the outliers by setting the three bits for visualization v as

$$v_{it} = 2^2 \hat{m}_{it}^{\text{cnv}} + 2^1 \hat{m}_{it}^{\text{meth}} + 2^0 \hat{m}_{it}^{\text{expr}} \quad (9)$$

where cnv indicates the CNV indicator, meth the methylation indicator, and expr the expression indicator. A heatmap was then generated for all genes $g \in \mathcal{G}$ for the gene set defined by the pathway $\mathcal{G}$, setting the least significant bit to red, the next bit to green, and the most significant bit to blue. To guarantee that all visualizations maintained the same color palette, we introduced a range bar at the bottom that included all possible values (i.e., 0 – 7) to overcome automatic rescaling of the heatmap in R. This visualization schema can be easily extended to more bits of information as additional measurements (e.g., mutation) become available.

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. As our goal in this study was to identify potential new treatable deregulated pathways in AML, our criteria for a pathway of interest were those significant in OGSA with approved therapeutics and not previously associated with AML. Other criteria would apply in different cases, naturally. We then used only the genes associated with the refined pathway list in TSP (e.g., those genes that define the gene set for this pathway in KEGG).

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. In many applications, the probability of $G_i = G_j$ is virtually zero, making the need to break the tie moot. However, for cases where $G_i = G_j$ is a probable outcome (i.e., scoring in pathology reports), ties can be broken by changing $G_i > G_j$ to $G_i \leq G_j$ [22].

The TSP is determined by finding the pair of genes that maximizes

$$\Delta_{ij} = \left| \frac{N_{TP}}{N_{\text{case}}} - \frac{N_{FP}}{N_{\text{control}}} \right|, \quad (10)$$

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 (here remission samples) and N_{control} controls (here diagnostic or relapse samples). 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 [9].

Here we used kTSP, as implemented in the R `ktspair` package [23]. We generated five TSPs in our training set for voting on the status of the samples.

Data and Preprocessing

We obtained data from the NCI TARGET initiative for pediatric AML. The data comprised 439 different tissue samples including diagnostic samples, remission samples from the same patients after completion of the initial treatment to attain remission, and relapse samples from some patients with a recurrence of AML. Measurements of mRNA levels (expression) were made with Affymetrix HuGene arrays on 201 diagnostic and 35 relapse samples only. Methylation measurements were made using Illumina HumanMethylation27 BeadChip arrays on 192 diagnostic and 192 matched remission samples. Measurements of copy number variation (CNV) were made using Affymetrix Genome-Wide Human SNP6 arrays on 188 diagnostic samples and 46 relapse samples only. All mapping to genes was done using build 18 (hg18) of the human genome using files from the UCSC Genome Browser [24].

Data types were preprocessed separately. Expression data was analyzed using the `oligo` package in R set to the core probes to get gene level summaries [25], with RMA applied for normalization [26]. Gene level expression estimates were obtained for 20, 165 named genes.

For methylation, β values representing the fraction of methylation were generated from U and M probe estimates in the Illumina FinalReport file using R [23]. Methylation estimates showing low variance across all samples were removed, leaving 19, 999 promoter methylation estimates associated with 11, 871 genes. Association between gene and probe was done by association of the probe with the gene immediately downstream of the probe.

For CNV, the arrays were processed using the `crmm` package in R [27]. CNV estimates were provided for all nonsynonymous probes, and the gene level copy number was estimated from the average value for all probes located within the gene transcription start and end sites. CNV estimates were obtained for 14, 493 named genes.

The integrated data set was constructed by combining gene level summaries. As there were genes with multiple methylation probes, we retained that probe that showed the highest variance across samples for a gene. We then eliminated all genes that were not present in each data type. The final data set comprised mRNA log2 expression estimates, fractional

methylation estimates on the highest variance probe per gene, and CNV estimates for genes measured on every platform, which provided 11, 158 genes.

Analysis of TARGET Methylation Data

We applied the OGSA and TSP methods to methylation data from the NCI TARGET initiative. From the methylation data of 192 diagnostic samples, 192 remission samples, and 46 relapse samples, 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 data set comprised 96 diagnostic samples and 96 remission samples from 96 patients, for a total of 192 methylation arrays. This karyotype-balanced set was chosen to avoid biasing the training set to any particular diagnostic subtype, as different karyotypes have different outcomes in AML [12]. Samples from the remaining 96 patients formed the test set (192 methylation arrays), and an additional set of remission and relapse samples was generated based on the 46 relapse samples and their matching remission samples (92 arrays, with remission samples overlapping with the previous sets).

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 in this analysis of methylation data only. Ranksum outlier analysis with an offset of $X_0 = 0.1$ was used to define outliers, and genes were ranked by the number of outliers. Significant pathways were determined by a Wilcoxon gene set test, and one pathway of interest was chosen based on its targetability. For this study, review of pathways was performed manually in light of the known biology of pediatric AML. This process could be automated in the future if databases that integrated pathway, disease, therapeutic, and clinical protocol information were available, as pathways significant in OGSA could be screened to determine if treatment for a pathway in a disease was already standard-of-care or in a clinical trial and if a therapeutic targeting a pathway member was already approved.

Five TSPs were generated from the probes associated with the genes assigned to the pathway using the ktspar 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 test data and to the relapse-remission data.

The 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).

Analysis of Integrated TARGET Data

As no remission data were available for mRNA or CNV molecular data types, we performed a different analysis on the integrated data. Using the training data samples, we identified the three year event status of all individuals and separated diagnostic sample data into event vs.

no event classes for outlier analysis. The data set comprised 72 diagnostic samples, with 37 classified as case = event and 35 as control = no event.

We performed outlier analysis using the three methods: Tibshirani-Hastie, Ranksum, and Corrected Ranksum with offsets $X_0 = \{1.0, 0.1, 0.5\}$ for expression, methylation, and CNV respectively. These offsets were motivated by typical limits applied in the field, as tissue heterogeneity often leads to compression of CNV signal, and methylation changes less than 10% are typically viewed as insignificant. We looked at outlier counts generated by the three methods, and we compared the genes with high numbers of outliers and the pathways deemed significant by the different methods using the VennDiagram package in R.

However, as the estimation of m could depend strongly on the offsets X_0 , we also explored the effect of varying X_0 in estimation of deregulated signaling and ranking of genes. We varied the thresholds as

$$X_0^{\text{expr}} = 0.0, 1.0, 2.0, 3.0, 4.0$$

$$X_0^{\text{meth}} = 0.0, 0.1, 0.2, 0.4, 0.75 \quad (11)$$

$$X_0^{\text{cnv}} = 0.0, 0.5, 1.0, 2.0$$

with a total of 30 different combinations tested. We looked at the correlation structure in the results in terms of outlier calls across genes and estimation of significant pathways.

As $\hat{m}_{it}^d$ provides an estimate of a specific molecular type of aberration (i.e., outlier) for each gene and each case sample, we also visualized one significant pathway from our analysis. In addition, to show that the uncorrected methods tend to call too many outliers, we visualized one pathway that was deemed insignificant with similar p -values in all methods.

Methylation Results

We applied our methods to the TARGET AML methylation data comprising 430 methylation 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 (Eqn. 7 with $D = 1$). These gene ranks were used to generate a set of significant pathways from the KEGG and Biocarta pathways using OGSA. We focused on one pathways from this set, the KEGG Hedgehog Signaling pathway, 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.

Significant Pathways from OGSA

Outlier analysis according to Eqn. 6 with an offset $X_0 = 0.1$ provided outlier ranks for all genes. The right-tail and left-tail rank lists were used in OGSA separately. The results of OGSA analysis of the KEGG and Biocarta pathway gene sets from the MSigDB database [28] are presented in Table 2. The p -values are Bonferroni corrected values from the Wilcoxon gene set test. All pathways with significant corrected 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 related to cancer (KEGG Basal Cell Carcinoma, Pathways in Cancer), 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). These processes are deregulated in most cancers and do not provide novel insights into 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 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 anti-apoptotic 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 Fig. 2). Hierarchical clustering demonstrated good separation of diagnostic samples (white bars on top) and remission samples (dark bars on top). In addition, good separation was also seen between matched relapse (lighter bars at top) and remission samples (see Fig. 2c).

In addition, it should be noted that the KEGG Basal Cell Carcinoma pathway may also be reflecting changes in Hedgehog gene methylation, as Basal Cell Carcinoma is often driven by Hedgehog signaling and Hedgehog inhibitors are used in treatment of advanced cases [29]. In adult AML, previous studies demonstrated hypermethylation of some members of the WNT signaling pathway [30], which interacts with the Hedgehog pathway.

KTSP Classifiers for the Hedgehog Pathway

In order to create a robust methylation signature for the Hedgehog pathway, we applied the KTSP algorithm to a subset of the Training data limited to promoter methylation levels of genes in the KEGG Hedgehog Signaling Pathway. We identified a set of 5 pairs that discriminate the diagnostic samples from the remission bone marrow samples (see Fig. 3,

where an X indicates a diagnostic sample and a filled square a remission 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 72 with a 95% confidence interval of [26, 236].

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 36 with a 95% confidence interval of [16, 92]. Interestingly, the application of the same signature to the Relapse data set was also predictive, now of relapse vs. remission, with $p = 1.4 \times 10^{-11}$, and an odds ratio of 30 with a 95% confidence interval of [9, 119]. This suggests that relapse in pediatric AML may be partially driven by recurrence of methylation changes in the promoters of Hedgehog signaling. Importantly, all tests show excellent Positive Predictive Values (93%, 86%, and 86% respectively), as is desirable for a test that could define treatment, since the vast majority of positive tests are related to positive pathway status.

Integrated Data Analysis

We also integrated the three data types available, expression, methylation, and copy number variation, for the 11,158 genes that were measured on all platforms. When multiple estimates were available for a single gene on any platform, we used the measurement showing the maximum variance across samples.

Outlier Counts and Gene Ranks

The integration required that biological behavior be followed, so right-tail outliers, representing overexpression and gene amplification in CNV, were summed with left-tail outliers in methylation, representing hypomethylation, as shown in Fig. 4. The method (here corrected rank sum) looked for genes that showed variation in tumors (red bars on right half of bar plots) that was outside the range in normals (blue bars on left half of bar plots) with minimum offsets of $X_0 = \{1.0, 0.1, 0.5\}$. For the top ranked gene NPAS2, the methylation changes determined 10 of the 11 outliers identified, with these outliers lying in the left edge of the red bars. Since the limit on expression is at least $\log_2$ of 1 and $\alpha = 0.05$, only a single expression outlier was called (red bar on right of NPAS Expr bar plot). As Figs. 5 a and b demonstrate, when the limits were removed, the outlier calls included very small methylation changes that are unlikely to play a role biologically (~ 0.05 for the Tibshirani-Hastie method and ~ 0.02 for the rank sum method). Looking at the range of values in Fig. 5c compared to these shows the effect of the limits, although it is also clear that the Tibshirani and Hastie method did perform somewhat better at finding outliers that appear strongly different from normals (Fig. 5a) than the uncorrected rank method (Fig. 5b).

The effect of applying limits to the rank outliers had two effects. First, it reduced the highest count of outliers for any gene (e.g., in Fig. 5 the total number of outliers across all three data types in the highest ranked gene were 22, 31, and 11, for the Tibshirani-Hastie method, rank method, and corrected rank method, respectively). Second, it also reduced the number of genes with a relatively high number of outliers, as shown in Fig. 6, where the number of genes with an outlier count greater than half the maximum in the method are compared for the three methods.

When carried forward to OGSA, the impact of the counting was very significant in the case of hypermethylation outliers integrated with underexpression and reduced copy number. In this case, both the rank sum method and the Tibshirani method identified 1 and 4 deregulated pathways at a Bonferroni corrected p -value of 0.05 respectively. The corrected rank sum method identified 12 pathways, including eight KEGG pathways clearly associated with cancer (focal adhesion, melanogenesis, pathways in cancer, regulation of the actin cytoskeleton, myeloid leukemia, glioma, melanoma, non-small cell lung cancer, and thyroid cancer). Only one of four pathways from the Tibshirani method (KEGG renal cell carcinoma) and none from the rank sum method were obviously associated with cancer, although cancer is complex and an association with other biological processes could well be valid. For instance, deregulation of the spliceosome, the top hit for the Tibshirani-Hastie method, is often associated with myeloid leukemias [31].

Robustness Analysis

In order to test the effect of different offset values X_0 on the corrected ranksum outlier method, we performed outlier analysis on 30 variations of offsets and looked at correlations between outlier counts and pathway p -values. As shown in Fig. 7, the correlations are quite high for almost all settings where $X_0^{\text{meth}}=0.1$. There is a second set of high correlations for $X_0^{\text{meth}}=0.2 - 0.4$ and $X_0^{\text{cnv}}=0.5$. High values or alternatively a 0 in an offset tended to show poor correlation with the more biologically plausible offset values.

As shown in the right hand panel in Fig. 7, the correlations were even stronger when the p -value estimates on pathways were the final goal. While high values of offsets again lowered correlations, excluding offsets of 0 or unreasonably high values led to strong correlations in estimates of pathway significance.

Tumor Specific Targets

When the $\hat{m}_{it}^d$ indicators are generated, they provide an indication of those specific genes that are potential targets for intervention, as well as the way in which those genes are aberrant relative to control samples. For instance, if $\hat{m}_{\text{MAGEA10, Pat25}}^{\text{meth}}=1$ in an analysis looking at left-tail outliers, then Patient 25 is predicted to have promoter hypomethylation of the MAGEA10 oncogene. Such analysis output provides information useful for understanding specific tumor etiology and, in cases of proteins with available drug interventions, potential information to aid treatment decisions.

Since pathways are the key deregulated biological entity in cancer, we visualized the outliers on a pathway specific basis. In Fig. 8, the corrected ranksum outlier analysis is summarized for one of the significant pathways in the overexpression (right-tail), hypomethylation (left-tail), and CNV amplification (right-tail) integrated analysis, the KEGG Extracellular Matrix (ECM) Receptor Interaction pathway. In this visualization with the bits set as outlined in Eqn. 9, a red element represents overexpression of the gene (column) in the tumor (row), green represents promoter hypomethylation, and blue represents increased copy number. RGB color mixing then permits visualization of cases where there are multiple outliers, such as the yellow elements indicating hypomethylation and overexpression.

The visualizations also permit easy comparison of the effects of the different outlier calling methods. Since the key statistic carried forward into gene set analysis is the rank of the gene in terms of outliers, overcalling of outliers may affect interpretation useful for patient-specific use much more dramatically than pathway calling. To demonstrate this point, we show in Fig. 9a pathway that was deemed highly insignificant in all three outlier count methods, the BioCarta AKAP95 pathway. In this case, the number and rank of outliers within the pathway must have been comparable to those outside the pathway to generate a very poor Wilcoxon p-value. However, as is clear from Figs. 9b and c, the uncorrected Ranksum method and the Tibshirani-Hastie method generated a large number of outlier calls. While these may technically be outliers, it is unlikely that this many calls would represent true biologically deregulation useful for interpreting cancer biology or medical intervention.

Conclusion

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 the relapse samples, suggesting both robustness of the marker and, potentially, that relapsed pediatric AML is driven partially by recovery of aberrant Hedgehog signaling. 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 [32]. 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.

While at a very early stage, the integration of data through outlier analysis shows great promise. The advantage of calling outliers within the individual data type reduces the complexity of analysis in the case where different data types have vastly different distributions and require different error models. For instance, methylation data are distributed as a beta function, expression data are typically treated as lognormal, copy number variation data tend to be sharply peaked with long tails, and mutation data may best be modeled as Poisson. Attempts to treat these data within a single mathematical framework are fraught with mapping issues, while summing outliers across different data types is trivially simple. Many questions remain as to the best way to identify outliers and to handle sums across samples. For instance, it would be useful to consider whether to count multiple hits to a single gene (e.g., hypermethylation and expression loss) as two outliers, as done

here, or to limit the counting to a single outlier per gene. In any case, it appears that incorporating biological plausibility will be very important.

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 statistical 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. MFO, JEF, MC, SM, and RJA were funded by the NIH/NCI U01 CA097452 National Childhood Cancer Foundation (TARGET). JEF is supported by the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000. YW was partially funded by NIDCR RC1DE020324. RJA also received support from the endowed King Fahd Chair in Pediatric Oncology and was in part supported by the Saint Baldrick's Foundation.

References

1. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000; 100(1):57–70. [PubMed: 10647931]
2. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011; 144(5):646–674. [PubMed: 21376230]
3. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz J, A L, Hartigan J, Smith DR, Strausberg RL, Marie SK, Shinjo SM, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. An integrated genomic analysis of human glioblastoma multiforme. *Science*. 2008; 321(5897):1807–1812. [PubMed: 18772396]
4. MacDonald JW, Ghosh D. Copa—cancer outlier profile analysis. *Bioinformatics*. 2006; 22(23):2950–2951. [PubMed: 16895932]
5. Naumov VA, Generozov EV, Zaharjevskaya NB, Matushkina DS, Larin AK, Chernyshov SV, Alekseev MV, Shelygin YA, Govorun VM. Genome-scale analysis of dna methylation in colorectal cancer using infinium humanmethylation450 beadchips. *Epigenetics : official journal of the DNA Methylation Society*. 2013; 8(9)
6. Waltering KK, Urbanucci A, Visakorpi T. Androgen receptor (ar) aberrations in castration-resistant prostate cancer. *Molecular and cellular endocrinology*. 2012; 360(1–2):38–43. [PubMed: 22245783]
7. Salk JJ, Fox EJ, Loeb LA. Mutational heterogeneity in human cancers: origin and consequences. *Annual review of pathology*. 2010; 5:51–75.
8. Smyth GK, Phipson B. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol*. 2010; 9:39.
9. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*. 2005 Oct; 21(20):3896–3904. [PubMed: 16105897]
10. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mrna comparison. *Statistical Applications in Genetics and Molecular Biology*. 2004; 3(1):19.
11. Price ND, Trent J, El-Naggar AK, Cogdell D, Taylor E, Hunt KK, Pollock RE, 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*. 2007; 104(9):3414–3419. [PubMed: 17360660]
12. Estey EH. Acute myeloid leukemia: 2013 update on risk-stratification and management. *American journal of hematology*. 2013; 88(4):318–327. [PubMed: 23526416]
 13. Liu KQ, Liu ZP, Hao JK, Chen L, Zhao XM. Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics*. 2012; 13:126. [PubMed: 22676414]
 14. Liu Y, Koyuturk M, Barnholtz-Sloan JS, Chance MR. Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. *BMC Syst Biol*. 2012; 6:65. [PubMed: 22694839]
 15. Ochs MF, Rink L, Tarn C, Mburu S, Taguchi T, Eisenberg B, Godwin AK. Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer research*. 2009; 69(23):9125–9132. [PubMed: 19903850]
 16. Kim YA, Wuchty S, Przytycka TM. Identifying causal genes and dysregulated pathways in complex diseases. *PLOS Computational Biology*. 2011; 7(3):e1001095. [PubMed: 21390271]
 17. Ulitsky I, Krishnamurthy A, Karp RM, Shamir R. Degas: de novo discovery of dysregulated pathways in human diseases. *PLoS One*. 2010; 5(10):e13367. [PubMed: 20976054]
 18. Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis. *Biostatistics*. 2007; 8(1):2–8. [PubMed: 16702229]
 19. Ghosh D. Discrete nonparametric algorithms for outlier detection with genomic data. *Journal of Biopharmaceutical Statistics*. 2010; 20(2):193–208. [PubMed: 20309754]
 20. Smyth, GK. *Limma: linear models for microarray data*. New York: Springer; 2005. p. 397–420.
 21. Kanehisa M, Goto S, Kawashima S, Nakaya A. The kegg databases at genomenet. *Nucleic Acids Res*. 2002; 30(1):42–46. [PubMed: 11752249]
 22. Xu L, Tan AC, Naiman DQ, Geman D, Winslow RL. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*. 2005; 21(20):3905–3911. [PubMed: 16131522]
 23. Gentleman RC, Carey VJ, Bates DM, 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 AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004; 5(10):R80. [PubMed: 15461798]
 24. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ. The ucsc genome browser database: update 2010. *Nucleic Acids Res*. 2010; 38(Database issue):D613–D619. [PubMed: 19906737]
 25. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*. 2010; 26(19):2363–2367. [PubMed: 20688976]
 26. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4(2):249–264. [PubMed: 12925520]
 27. Scharpf RB, Irizarry RA, Ritchie ME, Carvalho B, Ruczinski I. Using the r package crlmm for genotyping and copy number estimation. *Journal of statistical software*. 2011; 40(12):1–32. [PubMed: 22523482]
 28. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (msigdb) 3.0. *Bioinformatics*. 2011; 27(12):1739–1740. [PubMed: 21546393]
 29. Sekulic A, Mangold AR, Northfelt DW, LoRusso PM. Advanced basal cell carcinoma of the skin: targeting the hedgehog pathway. *Curr Opin Oncol*. 2013 May; 25(3):218–223. [PubMed: 23493193]
 30. Griffiths EA, Gore SD, Hooker C, McDevitt MA, Karp JE, Smith BD, Mohammad HP, Ye Y, Herman JG, Carraway HE. Acute myeloid leukemia is characterized by wnt pathway inhibitor promoter hypermethylation. *Leuk Lymphoma*. 2010 Sep; 51(9):1711–1719. [PubMed: 20795789]
 31. Ogawa S. Splicing factor mutations in myelodysplasia. *Int J Hematol*. 2012 Oct; 96(4):438–442. [PubMed: 23054646]

32. Knox SS, Ochs MF. Implications of systemic dysfunction for the etiology of malignancy. *Gene Regul Syst Bio*. 2013; 7:11–22.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. A flowchart of the analysis methods for performing outlier analysis on single or integrated data types, obtaining gene ranks, performing gene set analysis to identify significant pathways, and generating a biomarker using top-scoring pair for a significant pathway.

Figure 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 bars, dark indicates a remission sample, white a diagnostic sample, and light a relapse sample. Genes are in rows and patients in columns. Yellow (light color) indicates high methylation ($\beta \rightarrow 1$) and red (dark color) low methylation ($\beta \rightarrow 2$).

Figure 3.

The Five Top Scoring Pairs used to generate Table 3. X indicates a diagnostic sample and a filled square a remission sample.

Figure 4.

The highest ranked gene on integrated outlier analysis for the Corrected Rank Method with X_0 set to 1.0, 0.1, and 0.5 for log2 expression, methylation, and copy number data respectively. In each subplot, the bars on the left in blue represent non-event samples and the bars on the right in red represent event samples, i.e., the left samples are controls and the right are cases in the outlier analysis.

Figure 5.

The highest ranked genes on integrated outlier analysis shown for methylation for a) for the Tibshirani-Hastie method, b) for the uncorrected rank method ($X_0 = 0$), and c) for the corrected rank method. The bars are as in Fig. 4.

Figure 6.

The genes with a number of outliers greater than half the maximum is compared for each method in the case of a) hypermethylation outliers with underexpression and copy number loss and b) hypomethylation outlier with overexpression and copy number gain.

Figure 7.

The correlations of the outlier genes counts across all genes (left) and the gene set p-values across all pathways (right) for different threshold values. Threshold values are shown in the row and column labels in the order expression, methylation, CNV. As shown by the color key, high correlation values are lighter with the highest levels at the lower right, and low correlations are darker and cluster along the left and top edges.

The outlier calls are shown for the Extracellular Matrix Receptor pathway for the rank offset outlier method. The information content exceeds what can be presented in grayscale, so we describe here the color mapping. The bar at the bottom indicates possible outlier calls: red - expression, green - methylation, yellow - expression and methylation, blue - copy number, purple - expression and copy number, cyan - methylation and copy number, white - all three data type outliers. The individual bars indicate an outlier of the given type in a gene (column) and patient (row).

Figure 9.

A comparison of the outlier calls between the three methods, a) rank method with offsets, b) rank method without offsets, and c) Tibshirani-Hastie method, for an insignificant pathway with roughly the same p -value in each case. Colors are as in Fig. 8.

Table 1

Terms for Finding a 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

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	0.00005
KEGG ARRHYTH RT VENTR CARDIOMY-OPATHY ARVC	0.00008
KEGG BASAL CELL CARCINOMA	0.00027
KEGG CELL ADHESION MOLECULES CAMS	0.00027
KEGG FOCAL ADHESION	0.00199
KEGG CALCIUM SIGNALING PATHWAY	0.00216
KEGG PATHWAYS IN CANCER	0.01817
KEGG DILATED CARDIOMYOPATHY	0.02013
KEGG TYPE I DIABETES MELLITUS	0.03616
Left-Tail Outlier Results	p-Value
KEGG STEROID HORMONE BIOSYNTHESIS	< 0.00001
KEGG DRUG METABOLISM CYTOCHROME P450	< 0.00001
KEGG COMPLEMENT AND COAGULATION CASCADES	< 0.00001
KEGG RETINOL METABOLISM	< 0.00001
BIOCARTA COMP PATHWAY	< 0.00001
KEGG METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	< 0.00001
KEGG OLFACTORY TRANSDUCTION	< 0.00001
BIOCARTA CLASSIC PATHWAY	< 0.00001
KEGG TYROSINE METABOLISM	0.00008
BIOCARTA LECTIN PATHWAY	0.00048
KEGG LINOLEIC ACID METABOLISM	0.00160
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	0.00326
KEGG STARCH AND SUCROSE METABOLISM	0.00535
KEGG DRUG METABOLISM OTHER ENZYMES	0.01176
KEGG AUTOIMMUNE THYROID DISEASE	0.01739
KEGG ARACHIDONIC ACID METABOLISM	0.01744

Table 3

kTSP Hedgehog Only Classifier Performance

Training	Dx	Rm	Test	Dx	Rm	Relapse	RI	Rm
Call Dx	80	6		82	13	Call RI	38	6
Call Rm	16	90		14	83	Call Rm	8	40

Table 4

Significant Pathways in Integrated Data

Offset Rank Outlier Results	p-Value
KEGG PATHWAYS IN CANCER	< 0.00002
KEGG FOCAL ADHESION	0.00203
KEGG REGULATION OF THE ACTIN CY-TOSKELETON	0.00983
KEGG THYROID CANCER	0.01095
KEGG MELANOGENESIS	0.01099
KEGG GLIOMA	0.01317
KEGG MELANOMA	0.01493
KEGG NEUROTROPHIN SIGNALING PATH-WAY	0.03523
KEGG NON SMALL CELL LUNG CANCER	0.04021
KEGG CHRONIC MYELOID LEUKEMIA	0.04219
KEGG PROTEASOME	0.04233
BIOCARTA GH PATHWAY	0.04673
Rank Outlier Results	p-Value
KEGG NEUROACTIVE LIGAND RECEPTOR INTERACTION	0.01356
Tibshirani-Hastie Outlier Results	p-Value
KEGG SPLICEOSOME	< 0.00003
KEGG UBIQUITIN MEDIATED PROTEOLYSIS	0.00228
KEGG ENDOCYTOSIS	0.00254
KEGG RENAL CELL CARCINOMA	0.04501