INTERPRETING PERSONAL TRANSCRIPTOMES: PERSONALIZED MECHANISM-SCALE PROFILING OF RNA-SEQ DATA

ALAN PEREZ-RATHKE[†]

Department of Medicine, University of Illinois at Chicago
Chicago, IL 60612, USA
Email: perezrat@uic.edu

HAIQUAN LI[†]

Department of Medicine, University of Illinois at Chicago
Chicago, IL 60612, USA
Email: haiquan@uic.edu

YVES A. LUSSIER*,[†]

Departments of Medicine & Bioengineering, University of Illinois at Chicago
Chicago, IL 60612, USA
Email: ylussier@uic.edu

Despite thousands of reported studies unveiling gene-level signatures for complex diseases, few of these techniques work at the single-sample level with explicit underpinning of biological mechanisms. This presents both a critical dilemma in the field of personalized medicine as well as a plethora of opportunities for analysis of RNA-seq data. In this study, we hypothesize that the "Functional Analysis of Individual Microarray Expression" (FAIME) method we developed could be smoothly extended to RNA-seq data and unveil intrinsic underlying mechanism signatures across different scales of biological data for the same complex disease. Using publicly available RNA-seq data for gastric cancer, we confirmed the effectiveness of this method (i) to translate each sample transcriptome to pathway-scale scores, (ii) to predict deregulated pathways in gastric cancer against gold standards (FDR<5%, Precision=75%, Recall =92%), and (iii) to predict phenotypes in an independent dataset and expression platform (RNA-seq vs microarrays, Fisher Exact Test $p<10^{-6}$). Measuring at a single-sample level, FAIME could differentiate cancer samples from normal ones; furthermore, it achieved comparative performance in identifying differentially expressed pathways as compared to state-of-the-art cross-sample methods. These results motivate future work on mechanism-level biomarker discovery predictive of diagnoses, treatment, and therapy.

1. Introduction

Interpreting differentially expressed genes at the biological scale using enrichment statistics (Enrichment) or Gene-Set Analyses (GSA) has become routine for microarray and RNA-Seq studies. By design, these analyses require group assignment as well as derived mechanisms (e.g., Kyoto Encyclopedia of Genes and Genomes, i.e. KEGG pathways [1]) to reference differences of expression between these groups. While biologists are well served with such studies, evaluating individual patients in clinic necessitates single patient measures. Indeed, conventional single molecule biomarkers are popular because of their crisp thresholds that are interpretable as normal or abnormal. FDA-approved biomarkers are often required to reveal clinically interpretable biological mechanistic information useful in diagnosis of disease and prognosis of therapeutic response. While gene expression classifiers (signatures) have been shown as accurate predictors, they paradoxically are not comprised of "driver genes" (known mechanisms of diseases) or therapeutic response [2]. When developed using different datasets, there is poor genetic concordance between signatures. In contrast, we have shown mechanistic overlap at the protein interaction level between signatures predictive of clinical outcome in breast cancer [3] and in prostate cancer [4]. The lack of mechanistic underpinning prohibits in part the wide adoption and FDA approval of expression classifiers [5]. Indeed, MammaPrint® microarray [6] and of OncotypeDX [7] are both classifiers derived from mechanisms (wound healing signature from animal models, and curated breast cancer driver genes,).

Few genome-wide methods have been developed using gene-sets for imputing biological mechanisms (most have been for microarrays measuring RNA expression). In these studies, scoring mechanisms by the median or mean expression of their corresponding gene-set were shown to be capable of generating classifiers but at a lower accuracy than single-transcript RNA expression-level signatures [8, 9]. More accurate mechanism classifiers can be derived from methods comparing phenotypic group assignments between samples to identify principal components (PCA) [10, 11] or by the expression of key genes to represent the whole pathway such as in CORG [12] and LLR [13]. We developed "Functional Analysis of Individual Microarray Expression" (FAIME), a weighted rank method that can impute mechanism-scores on each expression array sample and eliminate the group assignment requirement [14]. We have shown FAIME's accuracy in generating classifiers predictive of outcome in independent expression array datasets of head and neck [14] and lung cancers [15]. We have also experimentally validated FAIME for predicting microRNA targets within cell lines and animal models [16]. We have additionally demonstrated that while the genetic overlap of RNA-level classifiers across three head and neck cancer datasets was ~3% at False Discovery Rate (FDR) <5%, more than 46%-61% of the FAIME-anchored KEGG pathways classifiers overlapped in the same datasets (FDR<5%) [14]. We have also demonstrated that FAIME can be employed on continuous phenotypes such as survival in cox-regression [12]. These studies [10-14] transcend those using conventional gene enrichment or gene set enrichment analyses (GSEA) that cannot provide individual measurements of mechanisms on a single sample and require comparison between multiple samples groups (in distinct categorical pheno-

types) to infer gene-set-level predictions. Recently, related work in mass spectrometry protein complexes (derived from interaction networks) were shown to be more accurate for designing classifiers than single proteins [17]. However, to our knowledge, no mechanism-level methodology has yet been designed specifically for interpreting individual RNA-sequencing samples. Such a methodology is a requirement to develop RNA-seq based, clinically predictive mechanism-level classifiers. To our knowledge, no method of mechanism imputation has been developed for RNA-seq at the single sample level.

We hypothesized that the FAIME weighted rank-based method we developed for expression arrays would be more accurate than the simpler 'median expression' and 'mean expression' methods. To confirm this for each method, we systematically compared the different false discovery rate thresholds for accuracy and for biological reproducibility across transcriptomic measurements using (i) proxy gold standards in the same datasets and (ii) validating in independent datasets (RNA-Seq vs array expression).

2. Methods

2.1. *Data preparation and databases*

All datasets were obtained from the Gene Expression Omnibus (GEO) [18]. To demonstrate the feasibility of the FAIME technique on RNA-seq data, the Asian gastric cancer dataset GSE36968 [19], consisting of 24 gastric cancers and 6 normal stomach samples, was used. GSE36968 was sequenced with Life Technologies SOLiD™ sequencing platform. This dataset was already in Reads Per Kilobase of exon model per Million mapped reads (RPKM) format [20]. Since RPKM is a widely accepted standard for RNA-seq normalization by biologists, no additional pre-processing was performed. To validate and show concordance among RNA-seq and microarray data, the Asian gastric cancer microarray dataset GSE13861 [21], consisting of 71 gastric cancer and 19 normal samples, was used. This dataset was already quantile normalized [22] and $\log_2$ transformed.

2.2. *Microarray platform annotation*

Microarray platform annotation was downloaded from the GEO website (http://www.ncbi.nlm.nih.gov/geo/) for the GSE13861 dataset using Illumina HumanWG-6 v3.0 expression beadchip.

2.3. *KEGG pathway annotations*

KEGG pathway annotations are embedded in Bioconductor database KEGG.db [23] version 2.7.1. The 229 KEGG pathways with more than 3 annotated genes are studied.

2.4. *FAIME pathway scoring of each sample*

From the methodologies in [1], to quantitatively assign a mechanism's "expression deregulation" via its gene members, whose expression is measured in RPKM, all expressed genes (set G) in each sample are sorted in a descending order according to their expression levels, and then, as shown in Eq. (1), an exponential decreasing weight (w) is as-

signed to the ordered genes. The resultant weighted expression values are used to prioritize relatively highly expressed genes as in the first step of Bioconductor package *OrderedList* [24, 25]. Specifically, let $r_{g,s}$ be the expression rank for each gene $g \in G$ in a sample *s*, let $|G|$ be the total number of distinct genes measured and the weight assigned to each gene per sample ($w_{g,s}$) is calculated as follows:

$$w_{g,s} = (r_{g,s}) \cdot (e^{\frac{r_{g,s} - |G|}{|G|}})$$ (1)

A Normalized Centroid (*NC*) is defined as the uni-dimensional average of the weighted expression values of a gene-set. Specifically, the sum of the weighted expression of gene element in a gene-set is normalized according to its cardinality. For every KEGG pathway, there is a gene-set *KEGGi* in which genes satisfy $g \in KEGG_i$ and a complement gene-set ($G/KEGG_i$) comprised of all available measured genes that are not annotated to this KEGG pathway. Thus we calculate the normalized centroid of each gene-set *KEGGi* in each sample *s* and that of its complement gene-set as follows:

$$NC(KEGG_{i,s}) = \frac{1}{|KEGG_i|} \sum_{g \in KEGG_i} (w_{g,s})$$ (2a)

$$NC(G/KEGG_{i,s}) = \frac{1}{|G/KEGG_i|} \sum_{g \in G/KEGG_i} (w_{g,s})$$ (2b)

$$where \quad G/KEGG_i = \{g : g \notin KEGG_i \cap g \in G\}$$

Furthermore, Eq. (3) calculates the Functional FAIME Score (*F* in equations) of each gene-set of a KEGG pathway in every sample as the difference between the normalized centroid of its gene-set and that of its complement gene-set. We define functional scores as functional biological mechanisms of the gene-set associated with a KEGG pathway in a given example.

$$F_{KEGG_{i,s}} = F(KEGG_{i,s}) = NC(KEGG_{i,s}) - NC(G/KEGG_{i,s})$$ (3)

Eq. (4) calculates for a sample *s*, the FAIME Profile "FP_s" defined as the set of all FAIME scores of sample *s*, $F_{KEGGi,s}$, assigned to every term.

$$FP_s = \{F_{KEGG_{1,s}}, \ldots, F_{KEGG_{i,s}}, \ldots, F_{KEGG_{n,s}}\}$$ (4)

where *n* is the total number of KEGG pathways.

In this way, patient-specific FAIME profiles of KEGG pathways are generated for each sample. Each sample has a continuous effective value for each category term which is the group difference between the genes annotated by the KEGG pathway and their individual complementary set of genes [16].

Calculations were performed using the latest FAIME R package which has been imrpoved to compute scores concurrently and allow for custom transformations (available: https://bitbucket.org/lussierlab/faime-opensource). Experiments were made with alternate transformations such as uniform-weighted rank and median selection, but we found that the original methodology performed the most consistently.

2.5. *Simpler methods for scoring each sample pathways*

To evaluate FAIME against alternative single-sample pathway scoring methods, we defined two unranked and two ranked methods. The unranked methods, *RPKM mean* and *RPKM median*, compute a sample's pathway score as either the mean or median of the RPKM values of the pathway's gene set respectively. Analogous rank-based methods, *Mean of Ranked RPKM* and *Median of Ranked RPKM*, first convert a sample's RPKM values into ranks and then score each pathway as the mean or median of the constitutive ranks respectively.

2.6. *Unsupervised hierarchical clustering (Figure 1)*

As seen in **Figure 1**, FAIME scores for all 229 KEGG pathways were used in generating the unsupervised hierarchical clustering of RNA-Seq dataset GSE36968. Similarly, other ranked methods (RPKM mean, RPKM median, mean of ranked RPKM and median of ranked RPKM) were employed for clustering as comparison. The clustered heat map was generated using the *heatmap* function of R with Ward's method as the distance criterion.

2.7. *Predicting deregulated pathways between two sets of samples using Wilcoxon parametric test (Figure 2&3, Table 1)*

In sections 2.4 and 2.5, we have described five methods (FAIME, RPKM mean, etc) that transform genome-wide RNA-seq or microarray-level measures of expression of a sample into pathway scores for this sample. Comparing samples of gastric cancer to normal gastric tissue, we calculate the deregulated pathways using the non-parametric Wilcoxon statistic and adjust for multiple comparisons using FDR. Thus, a set of deregulated pathways at different FDR thresholds can be imputed form the same dataset for each pathway scoring method. These can be compared to methods that calculate deregulated pathways directly from the gene-level expression such as GSEA and Enrichment studies (See section 2.8, ROC: Receiver Operating Characteristic).

2.8. *Evaluating pathway-scoring methods using ROC curves and proxy gold standards operating on the same RNA-seq dataset (Figure 2)*

Since it is unfeasible to biologically validate all predicted KEGG pathways, accuracy was determined using alternatively (i) GSEA [26] or (ii) conventional enrichment of differentially expressed genes (R package for SAM [27] analysis at FDR<5%) as proxy gold standards. At a given FDR, the set $positives_{GSEA}$ was calculated as the set of KEGG pathways found significantly differentially scored between cancer versus normal under GSEA (gene-set permutation); the set $positives_{FAIME}$ was calculated as the set of KEGG pathways found significantly differentially scored between cancer versus normal by running SAM [27] on the FAIME scores of each sample (Wilcoxon-statistic); the set $positives_{Enrichment}$ was calculated by first using SAM to identify significantly differentially expressed genes (Wilcoxon-statistic, fixed gene level FDR < 5%) and then performing hypergeometric enrichment on those genes for the KEGG pathways at the given FDR cutoff for pathways. Using GSEA as a proxy gold standard (**Figure 2, Panel A&B**), $positives_{GSEA}$

was fixed at FDR < 25% as recommended by the authors. Then, at various maximum FDRs ranging from 0% to 35%, the set of *true positives* for FAIME was calculated as $positives_{GSEA} \cap positives_{FAIME}$, the set of *false positives* as the set difference $positives_{FAIME} - positives_{GSEA}$, the set of *false negatives* as the set difference $positives_{GSEA} - positives_{FAIME}$, and the set of *true negatives* as the set difference $KEGG_{ALL}$ - (*true positives* $\cup$ *false positives* $\cup$ *false negatives*). With these values, we could then create a receiver-operating characteristic (ROC) curve for FAIME by plotting the *true positive rate* according to Eq. (A.1), versus the *false positive rate* according to Eq. (A.2). To compare with FAIME, a similar procedure was used to create the ROC curve for hypergeometric enrichment **(Figure 2, Panels C&D)**. To allow comparison of GSEA and FAIME, hypergeometric enrichment at FDR < 5% was instead used as a proxy gold standard and the corresponding ROC curves were created.

2.9. *Evaluating pathway-scoring methods in an independent dataset using concordance of prediction (Table1) and clustering (Figure 3)*

For each of the five pathway-scoring methods (see 2.4-2.6; FAIME, RKPM mean, etc), the R package for SAM [27] was successively used to prioritize pathways deregulated between gastric tumors and normal gastric tissue at FDR<2.5% and at FDR<5% in RNA-seq dataset GSE36968. The corresponding FAIME scores of those pathways in independent microarray dataset GSE13861 were then used as the basis for hierarchical clustering in **Figure 3** (R's *heatmap* function with Ward's method as the distance criterion). Similarly, differentially expressed pathways imputed from dataset GSE13861 at FDR 2.5% and 5% were used to hierarchically clustering samples in RNA-seq dataset GSE36968 and reported in **Table 1**. Furthermore, these analyses were successively conducted on the four other pathway-scoring methods: RPKM mean, RPKM median, mean of ranked RPKM, and median of ranked RPKM. The reciprocal study was conducted as well: prioritizing pathways for each method in the microarray studies and clustering the RNA-seq samples using the pathway scores of each RNA-seq sample corresponding to those prioritized pathways. Clustering accuracies of each method are reported in **Figure 3.** Further, an additional evaluation was conducted: the Fisher Exact Test (FET) and odds ratio of the concordance between the prioritized pathways derived independently over microarrays and RNA-seq are reported in **Table 1**.

3. Results and Discussion

To our knowledge, we present the first study of mechanism imputation at the single sample level for RNA-seq. This experiment differs from our previous ones in that we systematically also include as control intermediate geneset methods of computations such as mean, median, etc. In order to evaluate the feasibility to impute valid pathway scores at the individual sample level, we evaluated five distinct pathway-scoring methods in each of the following four experiments: (i) clinical phenotype clustering of individual RNA-seq samples by their pathways scores, (ii) concordance between pathways predicted at the RNA-seq single-sample level against those predicted at the cohort-wide level (such as in GSEA), (iii) the predictive power of prioritized pathways in one dataset as classification
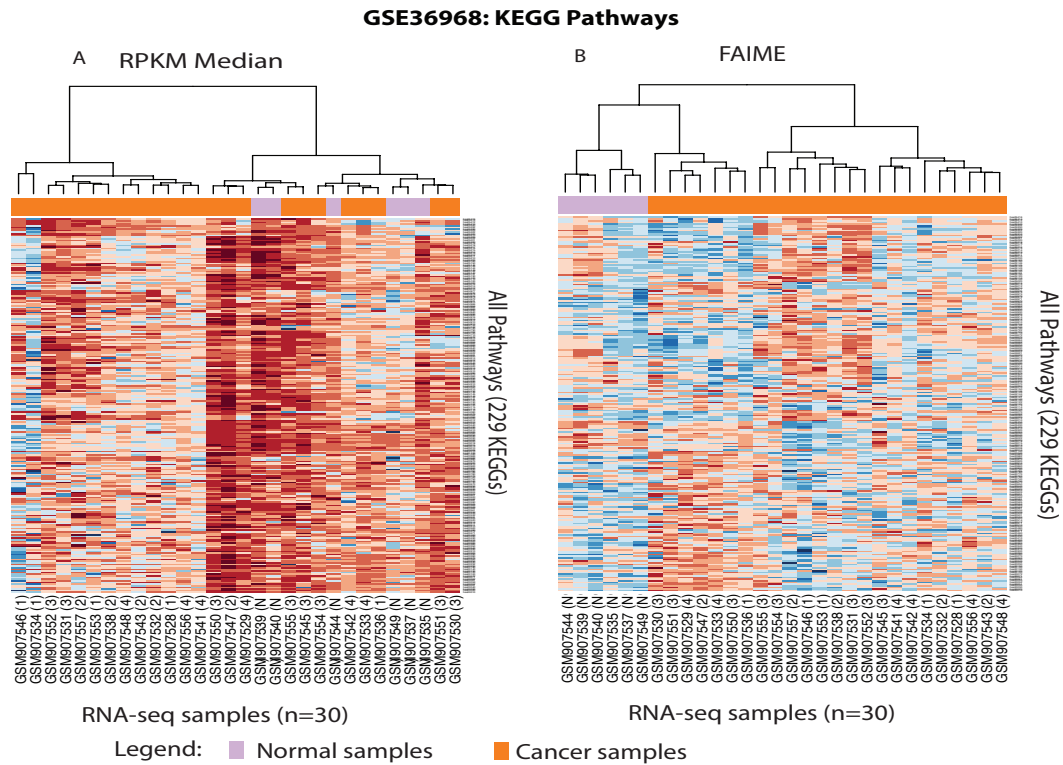
Figure 1. Unsupervised hierarchical clustering of all KEGG pathway-level scores imputed from RNA-seq RPKM of individual samples. *Panel A:* clustering RNA-seq dataset GSE36968 by "RPKM median" measure on individual sample ("RPKM means" - not shown - is equally inaccurate). *Panel B:* clustering RNA-seq dataset GSE36968 by FAIME scores imputed from individual samples (every other ranked-based method provided equally good clustering, not shown). This illustrates the pathway level clustering possible with pathway scoring at the single sample level (note: GSEA and Enrichment are not designed for this purpose). **Legend:** up-regulated pathways in cancer are blue and down-regulated ones are red. columns=30 samples; rows=229 KEGG pathways (formatted for reading at high magnification).

features for another dataset, and (iv) the concordance between pathway predictions conducted in two independent datasets (**Figure 1, Figure 2, Figure 3** and **Table 1,** respectively). In **Figure 1 Panel B**, FAIME scores for the entire KEGG ontology (229 pathways[1] were used to perform unsupervised hierarchical clustering. Using Ward's method [28] as the distance criterion, all normal samples were found within the same cluster, as were gastric cancer samples in RNA-seq dataset GSE36968. Other rank-based methods (mean of ranked RPKM, median of ranked RPKM) achieved similar clustering results but unranked methods (RPKM mean, RPKM median) failed to cluster accurately (**Figure 1, Panel A**). Note that cross-sample methods GSEA and Enrichment cannot work on single-sample level. Top panels in **Figure 2** demonstrate ROC curves for the KEGG pathways using GSE [26]as the proxy gold standard. For up-regulated pathways (**Figure 2 Panel A**), FAIME ROC performance compares favorably to hypergeometric enrichment. For down-regulated pathways (**Figure 2 Panel B**), FAIME and hypergeometric enrichment performed similarly. Bottom panels in **Figure 2** demonstrate ROC curves for the KEGG pathways using hypergeometric enrichment as the proxy gold standard. For both up-regulated (left) and down-regulated (right) pathways, FAIME ROC as the proxy gold
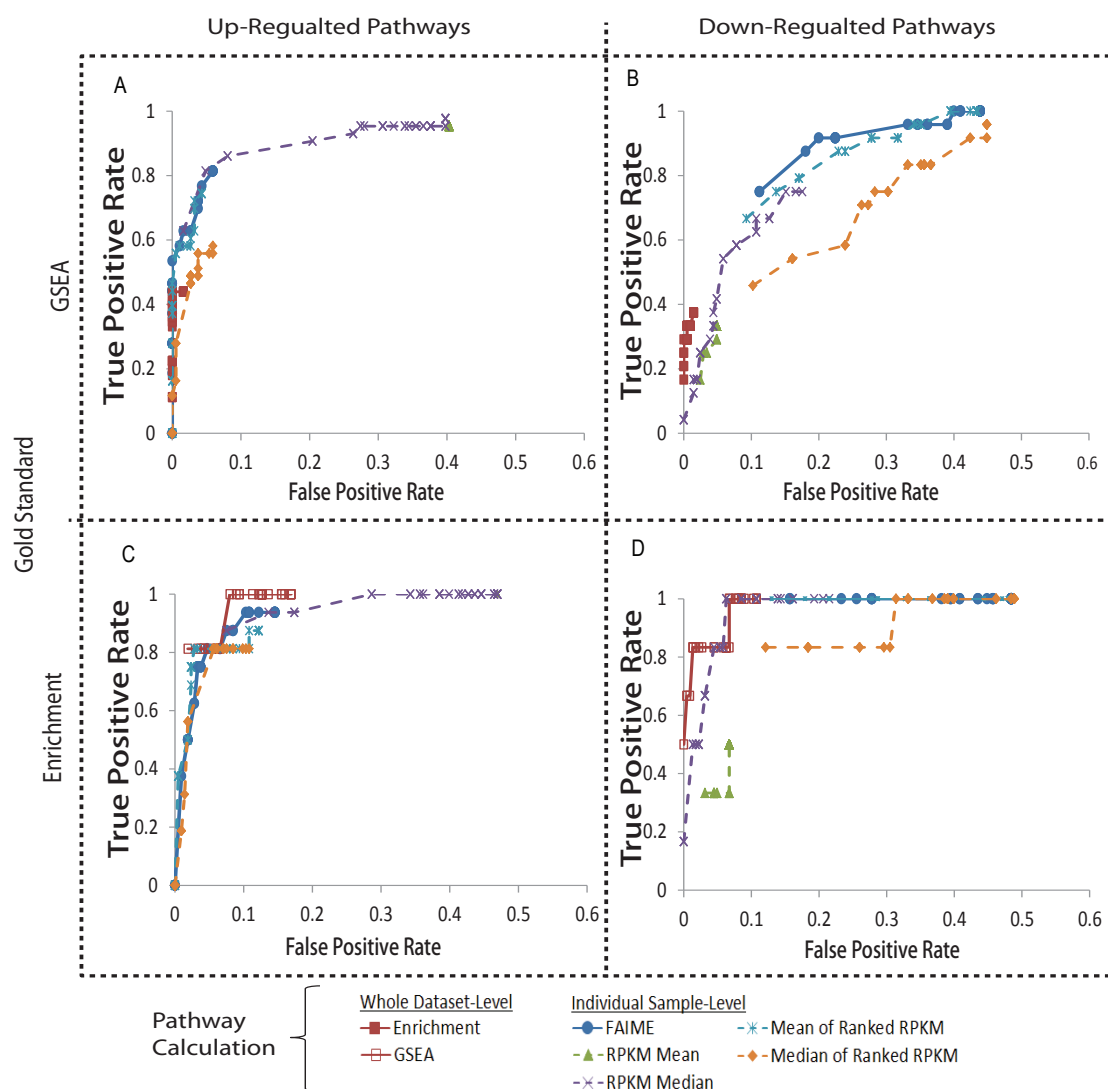
Figure 2: ROC curves of FAIME methods in identifying differentially expressed pathways as compared to GSEA, Enrichment, RPKM mean, RPKM median, mean of ranked RPKM and median of ranked RPKM. Panel A and B: ROC curves using differentially expressed pathways of GSEA as a proxy gold standard (FDR<25%). Panel C and D: ROC curves using differentially expressed pathways by Enrichment as a proxy gold standard (FDR<5%). Up- and down- regulated pathways vary at each accuracy threshold for each method and calculated is available at: http://lussierlab.org/publications/FAIME-rnaseq.

standard. For up-regulated pathways (**Figure 2 Panel A**), FAIME ROC performance is comparable to GSEA. We also compared the FAIME ROC performance with simpler, single-sample measures such as RPKM mean, RPMK median, mean of ranked RPKM and median of ranked RPKM (dashed lines) for both down-regulated pathways and up-regulated pathways, using either GSEA or enrichment method as benchmark. FAIME yields either superior or similar ROC performance as compared to these single-sample methods. The exception is the RPKM median method which surpasses ranked methods as well as RPKM mean.
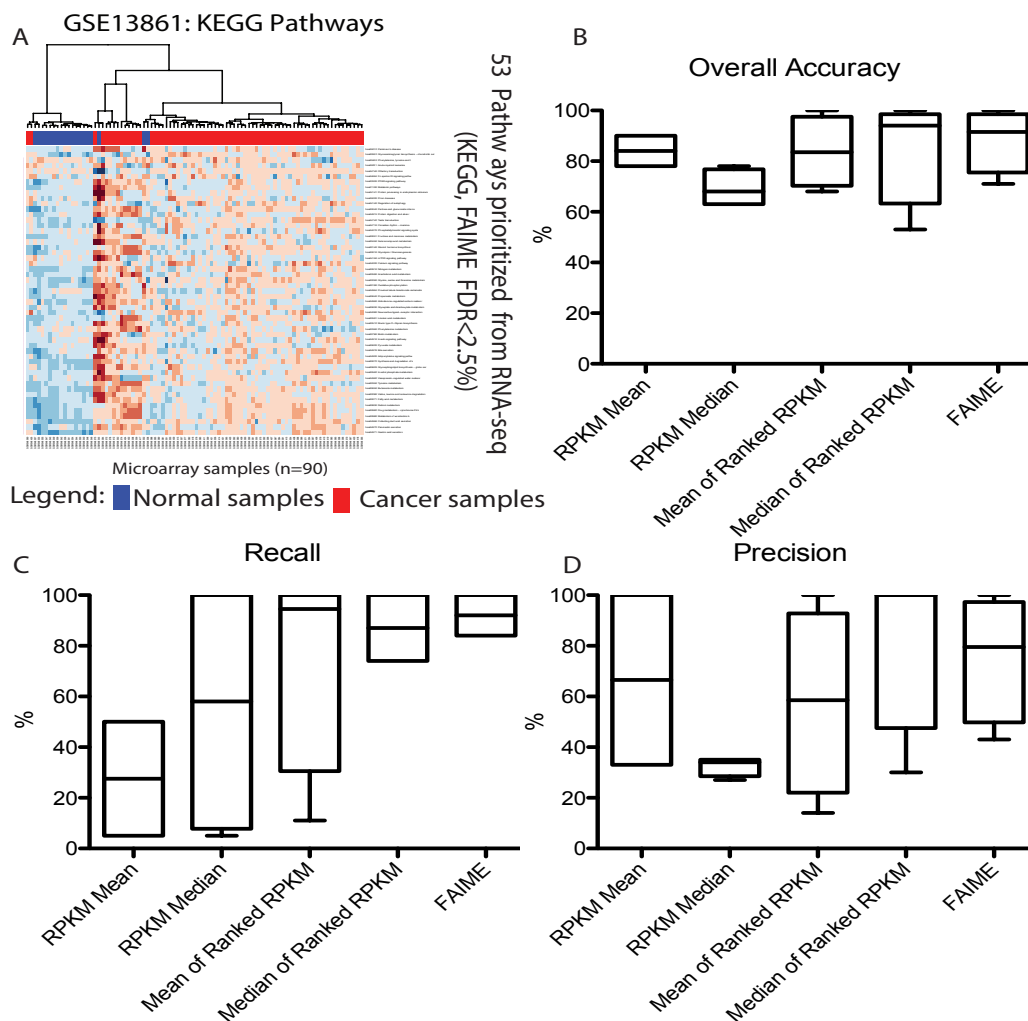
Figure 3: Unsupervised hierarchical clustering of gastric cancer datasets using sample-level scores of differentially expressed KEGG pathways learned from another independent dataset. Panel A: Clustering of microarray dataset GSE13861 by 53 significant different expressed FAIME pathways (FDR<0.025) learned from GSE36968 (large figure at http://lussierlab.org/publications/FAIME-rnaseq). **Panel B, C, D:** As described in **Methods (Section 2.6)**, deregulated pathways were prioritized in one dataset and their classification accuracies evaluated in an independent one (and vice-versa) producing accuracy scores reported here. Rank-based methods (mean of ranked RPKM, median of ranked RPKM, and FAIME) achieved overall better predictive performance across datasets as compared to unranked mean and median methods.

Figure 3a demonstrates hierarchical clustering of microarray dataset GSE13861 with 53 significantly differentially expressed FAIME features (FDR < 0.025) found in RNA- seq dataset GSE36968. 84 out of 90 (93.3%) samples are classified correctly. In a second set of experiments, reciprocal clustering of RNA-seq dataset GSE36968 using 122 and 140 differently expressed FAIME pathway features of microarray dataset GSE13861 (FDR <0.025 and FDR < 0.05 respectively). The overall accuracy, precision, and recall are shown in **Figures 3b**, **3d** and **3c** respectively. As shown from the three panels, RPKM median and RPKM mean methods achieved the worst results as compared to rank-basedmethods (mean of ranked RPKM, median or ranked RPKM, and FAIME).

Table 1. Pathway prediction concordance between the independent RNA-seq and microarrays datasets for each pathway-scoring method (Sub-table A). Sub-table B shows the stringent concordant subset of deregulated pathways prioritized by three techniques in both dataset (intersection): GSEA, Enrichment and FAIME that respectively predicted 29, 10 and 12 upregulated KEGG pathways and 21, 31 and 46 downregulated ones. Pathways known involved in gastric cancer are highlighted in blue (e.g. gemcitabine[5-FU], a pyrimidine analog, is a standard combination in treatment of gastric cancer). Detailed at: http://lussierlab.org/publications/FAIME-rnaseq

Sub-table A	Method	↓pathways* odds ratio	↓pathways* FET pvalue	↑pathways* odds ratio	↑pathways* FET pvalue
Dataset-wide metrics	GSEA	40	5×10^{-13}	60	$<2 \times 10^{-16}$
	Enrichment	14	6×10^{-3}	106	2×10^{-11}
Single Sample Metrics	Mean RPKM	12	2×10^{-4}	17	1×10^{-13}
	Median RPKM	14	5×10^{-3}	14	4×10^{-14}
	Mean of ranked RPKM	15	4×10^{-15}	*no overlap*	
	Median of Ranked RPKM	6	2×10^{-8}	*no overlap*	
	FAIME	19	$<2 \times 10^{-16}$	∞	1×10^{-6}

Type	KEGG ID	Sub-table B — KEGG Pathways
↑	04110	Cell cycle
↑	04115	P53 signaling pathway
↑	00240	Pyrimidine metabolism
↑	03040	Spliceosome
↑	03013	RNA transport
↑	03008	Ribosome biogenesis in eukaryotes
↓	00982	Drug metabolism – cytochrome P450
↓	00980	Metabolism of xenobiotics by cytochrome P450

Legend: * ↓↑respectively down- and up- regulated pathways in gastric cancer;
Odds ratio from the intersection between RNA-seq & array predictions; ∞: infinite (division by zero).

Evaluations conducted on the same dataset with proxy gold standards demonstrated that each method could produce modest to good accuracies - with the RPKM-mean method dominating. Paradoxically, the RPKM-mean was the worst method in term of recall and modest in terms of precision. This demonstrates that RPKM-mean is a volatile metric. In addition, the rank-based methods failed to identify up-regulated pathways in either GSE13861 or GSE36968 (Table 1). The FAIME method (which is a weighted rank-based method) achieved the most overall stable performance in reflecting the uniform underlying mechanisms across distinct types of datasets of the same gastric cancer diseases.

3.1. Future Studies and Limitations

While many studies have been completed in large RNA-seq datasets – they largely remain unavailable (either embargoed or simply not deposited in GEO). We are completing additional studies to corroborate the findings of this report in (i) other cancers, (ii) other diseases, and (iii) for predicting response to therapy. Identifying key genes in each pathway would merit to be evaluated in RNA-seq as well (e.g. CORG, [12]). Finally, other type of gene-sets beyond KEGG pathways and curated pathways should be considered. Co-expression modules derived from large scale studies of multiple disease conditions have provided insight in new biology and could be utilized as non-curated gene-sets. Protein complexes, that worked well in mass spectrometry [17], could also be utilized as gene-sets for pathway discovery in RNA-seq.

Further, we are exploring other pathway scoring approaches at the single-sample level that would conserve the inherent vectorial structure of pathway expression, without the

requirement of cross-sample analyses. We are also evaluating FAIME in a prospective clinical trial in predicting therapeutic response to recurrent head and neck cancer.

Additionally, FAIME exploits an exponential transformation algorithm that weights better highly expressed genes and thus rectifies (i) the saturation of microarray probes at high dynamic range and (ii) the high relative and absolute error rate (noise) on low expression measurements. Only the latter bias remains salient for RNA-seq. However, RPKM may not be the optimal metric for correcting biases of oversampling longer gene in next-gen seq. Moreover, most RNA-seq datasets are measured after reverse transcription on DNA-seq platforms, adding another potentially biased step to model. Thus, improving on mechanism-scoring methods for requires integrating modeling of new biases of specific RNA-seq platforms (e.g. adjustments for RNA fragment length that vary between platforms, gene length biases, reverse transcription, etc).

4. Conclusion

To demonstrate the feasibility of single-sample classification, we performed an entirely unsupervised hierarchical clustering of RNA-seq dataset GSE36968. This clustering does not rely on differentially expressed features found by a tool requiring multiple samples such as SAM [27] or GSEA. Instead, the FAIME scores for all KEGG pathways are used. Figure 1 demonstrates the success of this approach with 100% of normal samples being contained within the same cluster.

Accurate pathway-scoring techniques could conceivably be used as a single sample analysis mechanism whereby clinicians could establish a patient's pathway profile [14] as a diagnostic and prognostic utility. Identifying pathways with exceptionally high or low scores could also serve as a means to elucidate individualized drug targets. This could then allow for a personalized drug regimen based on transcriptomic analysis. However as shown with Mammaprint® and OncotypeDX®, the technologies adoption is complex and requires more than technical prowess.

Software availability

We provide a package allowing for high-throughput analyses of the five studied pathway-scoring methods on individual samples (https://bitbucket.org/lussierlab/faime-opensource).

Appendix

The *true positive rates* and *false positive rates* used in the ROC plots for FAIME, GSEA, and hypergeometric enrichment were calculated as follows:

$$true\ positive\ rate = \frac{|\ true\ positives\ |}{|\ true\ positives \cup false\ negatives\ |} \tag{A.1}$$

$$false\ positive\ rate = \frac{|\ false\ positives\ |}{|\ false\ positives \cup true\ negatives\ |} \tag{A.2}$$

References

1. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 1999. **27**(1): p. 29-34.
2. Massagué, J., *Sorting Out Breast-Cancer Gene Signatures.* New England Journal of Medicine, 2007. **356**(3): p. 294-297.

3. Chen, J., et al., *Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures.* Journal of Biomedical Informatics, 2010. **43**(3): p. 385-396.

4. Chen, J.L., et al., *Protein-network modeling of prostate cancer gene signatures reveals essential pathways in disease recurrence.* Journal of the American Medical Informatics Association, 2011. **18**(4): p. 392-402.

5. Kulasingam, V., M.P. Pavlou, and E.P. Diamandis, *Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer.* Nat Rev Cancer, 2010. **10**(5): p. 371-378.

6. Knauer, M., et al., *The predictive value of the 70-gene signature for adjuvant chemotherapy in early breast cancer.* Breast Cancer Research and Treatment, 2010. **120**(3): p. 655-661.

7. Paik, S., et al., *A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer.* New England Journal of Medicine, 2004. **351**(27): p. 2817-2826.

8. Guo, Z., et al., *Towards precise classification of cancers based on robust gene functional expression profiles.* BMC Bioinformatics, 2005. **6**(1): p. 58.

9. Abraham, G., et al., *Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context.* BMC Bioinformatics, 2010. **11**(1): p. 277.

10. Bild, A.H., et al., *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.* Nature, 2006. **439**(7074): p. 353-357.

11. Chen, X. and L. Wang, *Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer.* Journal of Computational Biology, 2009. **16**(2): p. 265-278.

12. Lee, E., et al., *Inferring Pathway Activity toward Precise Disease Classification.* PLoS Comput Biol, 2008. **4**(11): p. e1000217.

13. Su, J., B.-J. Yoon, and E.R. Dougherty, *Accurate and Reliable Cancer Classification Based on Probabilistic Inference of Pathway Activity.* PLoS ONE, 2009. **4**(12): p. e8161.

14. Yang, X., et al., *Single Sample Expression-Anchored Mechanisms Predict Survival in Head and Neck Cancer.* PLoS Comput Biol, 2012. **8**(1): p. e1002350.

15. Yang, X., et al. *Towards Mechanism Classifiers: Expression-anchored Gene Ontology Signature Predicts Clinical Outcome in Lung Adenocarcinoma Patients.* in *AMIA 2012 Annual Symposium.* 2012. Chicago.

16. Lee, Y., et al., *Network Modeling Identifies Molecular Functions Targeted by miR-204 to Suppress Head and Neck Tumor Metastasis.* PLoS Comput Biol, 2010. **6**(4): p. e1000730.

17. Goh, W.W.B., et al., *Proteomics Signature Profiling (PSP): A Novel Contextualization Approach for Cancer Proteomics.* Journal of Proteome Research, 2012. **11**(3): p. 1571-1581.

18. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Research, 2002. **30**(1): p. 207-210.

19. Kim, Y.H., et al., *AMPKα Modulation in Cancer Progression: Multilayer Integrative Analysis of the Whole Transcriptome in Asian Gastric Cancer.* Cancer Research, 2012. **72**(10): p. 2512-2521.

20. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Meth, 2008. **5**(7): p. 621-628.

21. Cho, J.Y., et al., *Gene Expression Signature–Based Prognostic Risk Score in Gastric Cancer.* Clinical Cancer Research, 2011. **17**(7): p. 1850-1857.

22. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-193.

23. Marc Carlson, S.F., Herve Pages and Nianhua Li *KEGG.db: A set of annotation maps for KEGG.*

24. Lottaz, C., et al., *OrderedList—a bioconductor package for detecting similarity in ordered gene lists.* Bioinformatics, 2006. **22**(18): p. 2315-2316.

25. YANG, X., et al., *SIMILARITIES OF ORDERED GENE LISTS.* Journal of Bioinformatics and Computational Biology, 2006. **04**(03): p. 693-708.

26. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.

27. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences, 2001. **98**(9): p. 5116-5121.

28. Ward, J.H., Jr., *Hierarchical Grouping to Optimize an Objective Function.* Journal of the American Statistical Association, 1963. **58**(301): p. 236-244.