Biological function annotation of microarray data using Pathway Index

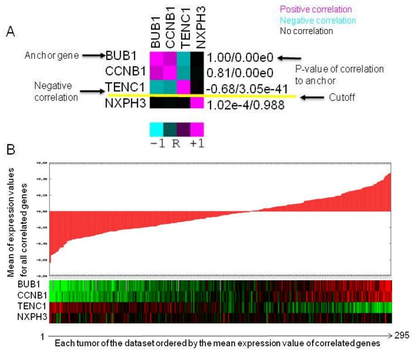
Bin Feng, Kathryn Sun, Richard Nicoletti, James Abraham, and Murray O. Robinson*

AVEO Pharmaceuticals, Inc., Cambridge, MA, USA

Abstract

- **Motivation:** Although whole-genome microarray analysis has become a routine tool in biomedical research, extracting meaningful information remains a challenge. We describe a method, Pathway Index, to assess biological information in microarray data that relies on the observation that genes common to a biological function often exhibit correlated expression within a microarray dataset. Pathway Index uses gene sets representing a biological process of interest, determining the subset of genes that exhibit significantly correlated/anti-correlated expression, followed by the mean expression of the set of correlated genes.
- **Results:** Using Pathway Index, we discovered the relationships between two distinct prognostic breast cancer signatures, MammaPrint and Oncotype DX, and basal subtypes within a breast tumor dataset. Our analysis indicated that both MammaPrint and Oncotype DX are good prognostic biomarkers, and both biomarkers correlated with basal signature score. However, we also demonstrated that both MammaPrint and Oncotype DX predicted patient outcome in a non-basal type cohort, which suggested that in addition to identify basal subtype as bad outcome group, both MammaPrint and Oncotype DX were good prognostic biomarkers in non-basal type cohort. We then established a biological correlation between a previously identified chromosomal instability signature and a proliferation signature in human cancer. We then annotated a renal cancer dataset with Pathway Index scores of 183 canonical pathways, to identify hypoxia deregulation in Renal Cell Carcinoma samples. We also applied this approach to analyzing genes along chromosome 7 and identified amplification of the EGFR locus using only transcriptome microarray data. Thus, Pathway Index provides a new tool to annotate microarray data with biological pathway information.

Figure 1.



Method

The Pathway index approach is a general unsupervised pathway analysis method for microarray data analysis. Given a predefined set of genes *S* (e.g., genes in the same pathway), the Pathway Index methodology determines the relative transcriptional flux (coordinated changes in genes common to a pathway across a given dataset of N individual samples) of the correlated subset of probes in *S*. Similar to any statistical analysis of microarray data, this approach assumes that the biological state of the queried pathway exhibits variation across the assayed dataset. Pathway Index uses the mean expression values of correlated genes in each pathway to represent the pathway status. Determining a Pathway Index can be divided into three steps:

Step 1: Identification of the anchor gene.
Step 2: Rank genes in correlation matrix.
Step 3: Inversion of the expression values of negatively correlated genes.
Step 4: Averaging expression value of correlated genes in the set *S*.

Figure 1. (A) Correlation matrix for one example of four-gene proliferation signature. BUB1 correlated the best with the other three genes and was identified as the anchor gene for the signature. The correlation matrix was ranked by p-value of correlation to the anchor gene. The yellow line was the 0.01 p-value cutoff for correlated expressed genes within the signature. (B) Pathway Index scores were calculated by the average expression of correlated expressed genes in the signature. The expression values of TENC1 were flipped because TENC1 was negatively correlated with the anchor gene BUB1. Upper panel showed the four-gene proliferation signature Pathway Index scores of each of the 295 samples in the NKI breast tumor dataset from low to high. Bottom panel showed the expression value heatmap of these four genes aligned the same order as the upper panel.

Figure 3. Pathway Index scores of 70-gene MammaPrint signature correlated with those of 16-gene Oncotype DX signature. Both Pathway Index scores of MammaPrint and Oncotype DX correlated with Basal signature Pathway Index scores. The table showed the Pearson correlation coefficient between Pathway Index scores of basal signature, MammaPrint signature and Oncotype DX signature in NKI breast dataset.

Results

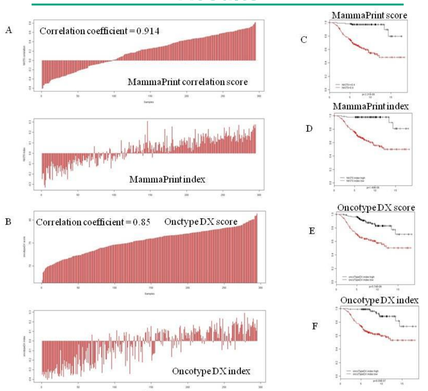


Figure 2. (A) Pathway Index scores of MammaPrint 70-gene signature correlated with MammaPrint correlation coefficient scores in NKI dataset. Upper panel showed the 70-gene MammaPrint correlation coefficient scores from low to high. Lower panel showed the Pathway Index scores of 70-gene MammaPrint signature aligned the same order as the upper panel. (B) Pathway Index scores of Oncotype DX 16-genes signature correlated with Oncotype DX scores in NKI dataset. Upper panel showed the 16-gene Oncotype DX scores from low to high. Lower panel showed the Pathway Index scores of 16-gene Oncotype DX signature aligned the same order as the upper panel. (C) MammaPrint correlation coefficient scores predicted survival in NKI dataset. (D) Pathway Index scores of 70-gene MammaPrint signature predicted survival in NKI dataset. E. Oncotype DX scores predicted survival in NKI dataset. D. Pathway Index scores of 16-gene Oncotype DX signature predicted survival in NKI dataset.

Figure 3.

Correlation coefficient	Basal signature	MammaPrint	Oncotype DX
Basal signature	1	0.524	0.593
MammaPrint	0.524	1	0.803
Oncotype DX	0.593	0.803	1

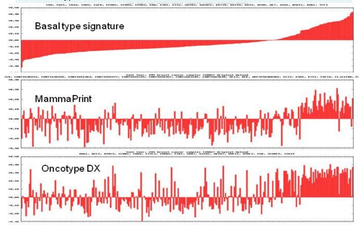


Figure 4. (A) MammaPrint predicted survival in non-basal subset of NKI breast dataset. (B) Pathway Index scores of 70-gene MammaPrint signature predicted survival in non-basal subset of NKI breast dataset. (C) Oncotype DX predicted survival in non-basal subset of NKI breast dataset. (D) Pathway Index scores of 16-gene Oncotype DX signature predicted survival in non-basal subset of NKI breast dataset.

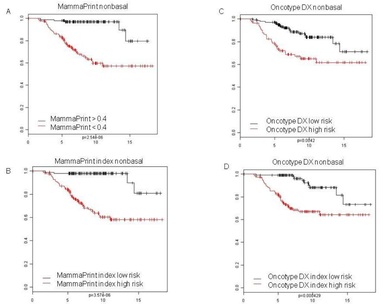
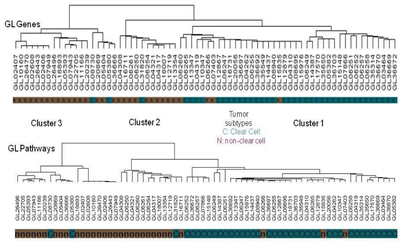


Figure 5. Comparison between hierarchical clustering by genes and hierarchical clustering by Pathway Index scores of canonical pathways in the GeneLogic Kidney tumor dataset.



	Genes	Pathways	overlap	Percentage
cluster 1	40	39	39	97.5%
cluster 2	11	12	11	100%
cluster 3	19	19	19	100%

Figure 6. Pathway Index scores of proliferation signature, CIN25 and the 21-gene non-proliferation CIN25 signature correlated with each other in NKI breast tumor dataset.

Correlation coefficient	Proliferation	CIN25	CIN25-no.proliferation
Proliferation	1	0.966	0.960
CIN25	0.966	1	0.997
CIN25-no.proliferation	0.960	0.997	1

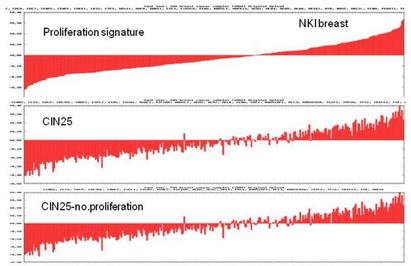


Figure 7. (A) CIN25 predicted survival in NKI breast tumor dataset. (B) The proliferation signature predicted survival in NKI breast tumor dataset.

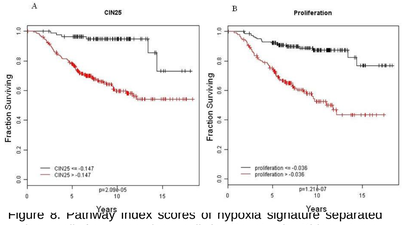


Figure 8. Pathway Index scores of hypoxia signature separated clear cell from non-clear cell in GeneLogic Kidney tumor dataset. Pathway Index scores of hypoxia signature were calculated and ordered from low to high.

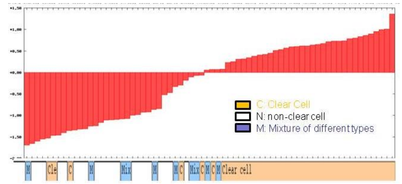
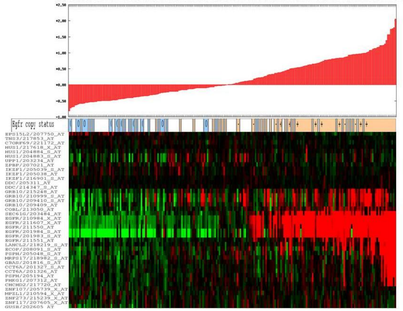


Figure 9. Pathway Index scores of EGFR locus co-linear signature based on expression microarray predicted EGFR copy number amplification status in TCGA GBM dataset. EGFR copy number status was based on array-CGH data.



Conclusions

- Pathway Index calculates a score for each pathway to represent the activation status of the pathway.
- A second feature of the Pathway Index approach is shown by demonstrating that calculated Pathway Index scoresof 183 canonical pathways (Ingenuity Inc., Redwood City, CA) preserved whole genome profile information by showing similar hierarchical clusters with either Pathway Index scores or individual genes. Unlike the individual gene clusters however, the pathway index calculations reduce the dimensionality of the data, while providing additional information on the state of the canonical pathways.
- Pathway Index can also be used to analyze gene signatures.
- Another application of Pathway Index is estimating copy number based on mRNA microarray data.

References

1. van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) Gene expression profiling predicts clinical outcome of breast cancer, Nature, 415, 530-536.
2. van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H. and Bernards, R. (2002) A gene-expression signature as a predictor of survival in breast cancer, N Engl J Med, 347, 1999-2009.

Acknowledgments

Study supported by AVEO Pharmaceuticals, Inc., Cambridge, MA.