Differential Diagnosis of Hematologic and Solid Tumors Using Targeted Transcriptome and Artificial Intelligence

Hong Zhang1, Muhammad Asif Qureshi2, Mohsin Wahid2, Ahmad Charifa1, Amir Ehsan3, Andrew IP4, Ivan De Dios1, Wanlong Ma1, James McCloskey4, Michele Donato4, David Siegel4, Martin Gutierrez4, Andrew Pecora4, Andre Goy4, Maher Albitar1.

1Genomic Testing Cooperative, 2Dow University of Health Sciences Karachi, 3CorePath Laboratories, 4 John Theurer Cancer Center, Hackensack Meridian Health.

Background:

RNA profiling of cancer has been shown to be highly informative providing information on the tumor, microenvironment, and immune response. Using next generation sequencing (NGS) in analyzing RNA is making RNA profiling a reliable clinical tool and approach for the discovery of biomarkers, characterizing the biology of the cancer and prediction of outcome and response to various therapeutic approaches. RNA sequencing and quantifying expression using NGS is significantly more reliable and reproducible as compared with old methodologies such as microarrays or PCR-based RNA quantification. Targeted RNA sequencing of various tissue samples allows us to focus on relevant oncogenic markers and allows us to sequence at deep level for better quantification of low level expressor genes that might be very relevant as major regulator of the complex biology of cells.

We explored the potential of using targeted transcriptome and artificial intelligence in the differential diagnosis and classification of various hematologic and solid tumors.

Methods:

For machine learning, we first selected genes that distinguish between two classes using standard Naïve Bayesian classifier on each gene with k-fold cross validation. After selecting individual genes, we used Naïve Bayesian classifier to distinguish between diagnostic classes using multiple selected genes using both confidence and P values. However, since Naïve Bayesian classifier suffers from severe numerical underflow problem when the dimension of data is high, we developed the Geometric Mean Naïve Bayesian (GMNB) classifier that eliminates the underflow problem by applying a multiplicative positive increasing function to the likelihood. The Geometric Mean Naïve Bayesian (GMNB) classifier was also used in classifying each sample against multiple diagnostic classes.

Conclusions

- Targeted transcriptome when used with machine learning is highly reliable in distinguishing between two diagnostic classes
- Artificial intelligence and targeted transcriptome can aid in diagnosis-making decision by ranking diagnostic probabilities between 47 different hematologic and solid tumors.

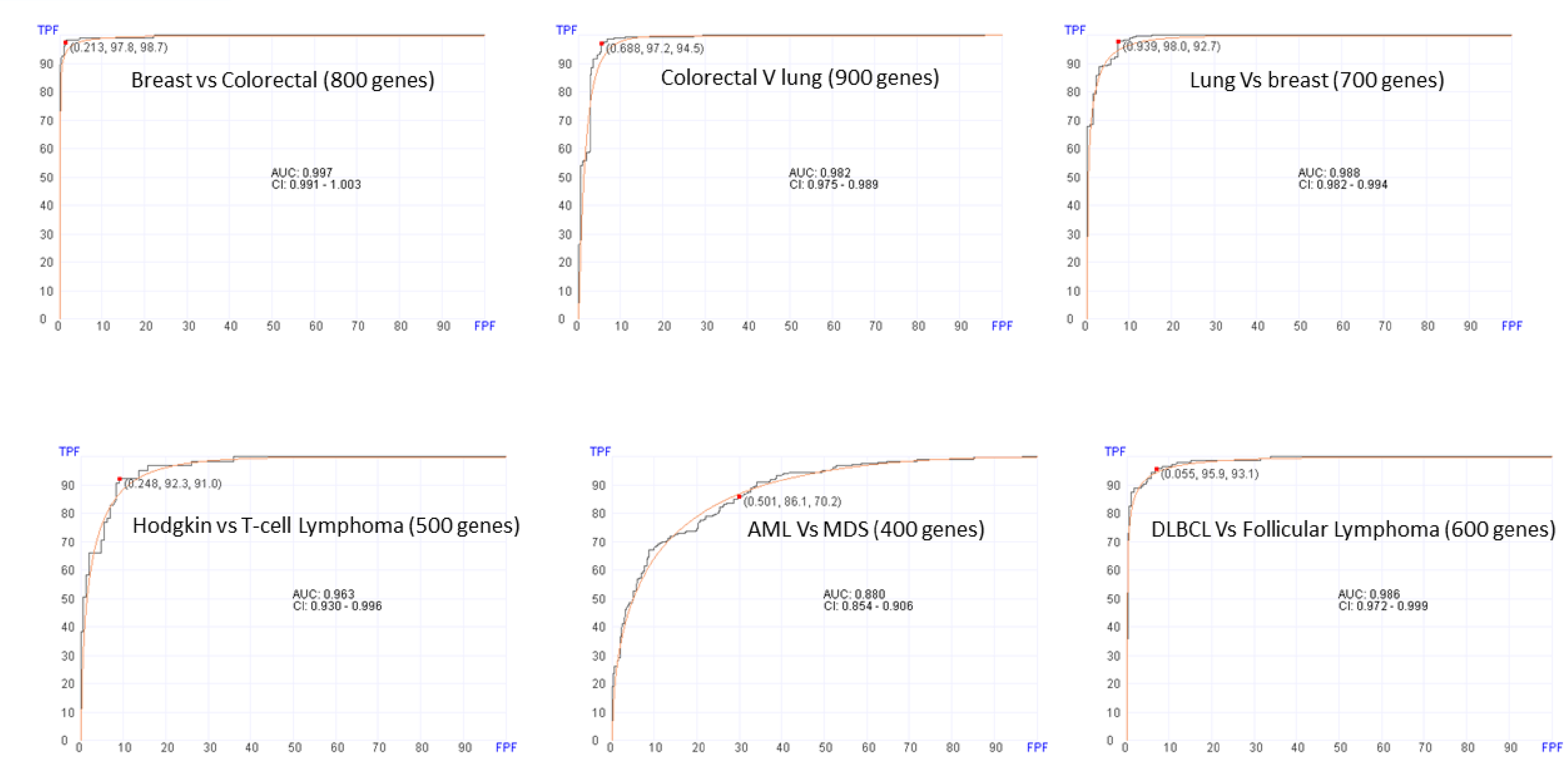
Results

Differential diagnosis between 47 different diagnostic classes with ranking

Diagnosis	No. of cases	No. of accurately diagnosis as 1st choice (positive percent agreement)	Positive predictive value (PPV)	No. of accurately diagnosis as 2nd choice	Positive percent agreement (PPA) by 1st and 2nd choices
ALL	26	26 (100%)	84%	0 (0)	100%
Colorectal	101	83 (82%)	79%	4 (4%)	86%
Brain	16	12 (75%)	75%	0 (0%)	75%
Lung	201	177 (88%)	73%	7 (3%)	91%
DLBCL	149	127 (85%)	73%	8 (5%)	91%
breast	31	25 (81%)	71%	2 (6%)	87%
CLL	61	44 (72%)	69%	5 (8%)	80%
Endometrial	31	21 (68%)	66%	3 (10%)	78%
MM	31	22 (71%)	65%	0 (0%)	71%
Ovarian	41	29 (71%)	63%	6 (15%)	85%
Pancreas	31	19 (61%)	58%	5 (16%)	77%
follicular	36	26 (72%)	53%	5 (14%)	86%
Mantle	31	18 (58%)	50%	3 (10%)	68%
Sarcoma	40	26 (65%)	45%	1 (3%)	68%
Hodgkin	26	16 (62%)	41%	9 (35%)	97%
Normal	201	92 (46%)	37%	39 (19%)	65%
AML	120	106 (88%)	35%	6 (5%)	93%
T-cell	41	21 (51%)	34%	8 (20%)	71%
Marginal	26	8 (31)	26%	4 (15%)	46%
MDS	101	19 (19%)	13%	47 (47%)	65%
MPN	26	3 (12%)	9%	3 (12%)	23%
CMML	31	2 (6%)	4%	2 (6%)	13%
CML	17	0 (0%)	0%	1 (6%)	6%

High accuracy in the differential diagnosis between two diagnostic classes

Two classes	AUC (95% CI)	Sensitivity (%)	Specificity (%)	No. of genes	Leave one out AUC (95%CI)
Normal Vs AML	0.9764 (0.954-0.974)	90.9	93.2	100	0.945 (0.933-0.957)
Normal vs ALL	0.981 (0.973-0.989)	95.1	95.5	200	0.977 (0.968-0.985)
Normal vs CLL	0.997 (0.994-0.999)	96.4	98.8	100	0.980 (0.973-0.988)
Normal vs Mantle	0.992 (0.987-0.997)	95.1	97.8	100	0.969 (0.959-0.980)
Normal vs MDS	0.831 (0.801-0.861)	78.1	75.3	400	0.826 (0.796-0.856)
Normal vs MPN	0.923 (0.884-0.962)	90.9	82.3	400	0.903 (0.860-0.946)
MDS Vs MPN	0.884 (0.837-0.931)	90.9	70.8	500	0.806 (0.748-0.864)
AML vs MDS	0.880 (0.854-0.906)	86.1	70.2	400	0.864 (0.837-0.892)
CLL vs Mantle	0.986 (0.968-1.000)	94.6	95.2	10	0.986 (0.968-1.00)
Marginal vs CLL	0.984 (0.964-1.00)	98.7	91	25	0.86.64 (0.809-0.920)
Marginal vs follicular	0.946 (0.917-0.974)	91	93.4	550	0.942 (0.912-0.971)
Hodgkin vs Normal LN	0.990 (0.972-1.00)	95.4	100	100	1.00 (1.00-1.00)
Hodgkin Vs T-cell lymphoma	0.963 (0.930-0.996)	92.3	91	500	0.902 (0.850-0.954)
Hodgkin Vs DLBCL	0.975 (0.948-1.00)	96.9	95.3	500	0.965 (0.934-0.997)
DLBCL vs Follicular	0.986 (0.972-0.999)	95.9	93.1	600	0.975 (0.957-0.993)
DLBCL vs T-cell lymphoma	0.967 (-.946-0.988)	91.7	89.8	600	0.942 (0.915-0.969)
Lung vs Colorectal	0.982 (0.975-0.989)	97.2	94.5	900	0.977 (0.969-0.985)
Lung Vs breast	0.988 (0.982-0.994)	98	92.7	700	0.988 (0.982-0.994)
Breast vs ovarian	0.994 (0.984-1.00)	100	94.2	700	0.989 (0.976-1.00)
Ovarian vs endometrial	0.959 (0.933-0.984)	92.9	91.2	600	0.853 (0.803-0.902)
Breast Vs Colorectal	0.997 (0.991-1.00)	97.8	98.7	800	0.987 (0.973-1.00)
Pancreas Vs colorectal	0.989 (0.980-0.997)	94.5	95.8	550	0.971 (0.956-0.985)
pancreas Vs esophageal	0.999 (0.990-1.00)	97.1	98.9	550	0.960 (0.914-1.00)
Ovarian vs lung	0.994 (0.984-1.00)	97.6	96.6	600	1.00 (0.997-1.00)
Lung vs DLBCL	0.996 (0.992-0.999)	97.2	97.3	800	0.988 (0.983-0.993)
Sarcoma Vs Ovarian	0.995 (0.986-1.00)	99.2	95.7	300	1.00 (0.997-1.00)
Sarcoma vs GIST	1.00 (0.997-1.00)	99.3	100	300	1.00 (0.997-1.00)



Future Directions for Research:

We are adding mutations profile and clinical data to the machine learning algorithm and exploring the reliability of this approach in diagnosis, predicting prognosis and response to therapy.