CERTIS™ ONCOLOGY INTELLIGENCE

Gene Expression-Based Machine Learning Classifier to Predict and Validate Cancer Type in PDX Models

Warren Andrews, PhD[1]; Jonathan Nakashima, PhD[1]; Long Do, PhD[1]

AACR 2023 ABSTRACT NUMBER 4299/30

Patient-derived xenograft (PDX) models are increasingly utilized in translational research and drug development.[2] Characterizing the genomic features of PDX is essential to establishing reliable models for cancer research.[3] Despite great interest, problems remain in PDX tumor data banks, including improper cancer-type diagnosis and sample mix-ups. To improve annotation and quality of PDX models, Certis developed a machine learning model trained on gene expression data from the Cancer Genome Atlas (TCGA). Certis then applied the model to corresponding data collected from nearly 300 Certis PDX models as well as the National Cancer Institute's (NCI) Patient-Derived Models Repository (PDMR). The model shows high precision and variable recall and provides a fast and accurate method for cancer-type diagnosis.

METHODS

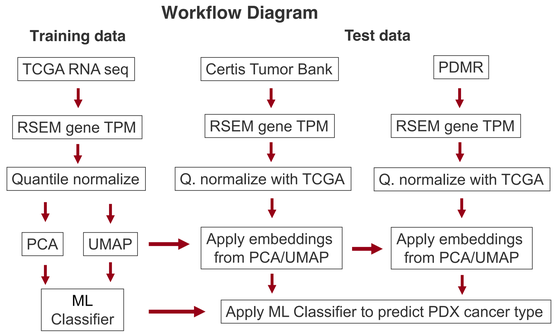
Overview: Because the total number of genes (~50k) is much larger than the number of samples in TCGA (~10k) or most PDX tumor banks (~100 to 1k), training an ML model on genes directly is suboptimal due to the parameter space being too broad. To solve this problem, Certis employed two dimension reduction approaches in parallel: UMAP[4] and Principal Component Analysis (PCA). Certis found these two approaches to have similar performance and to be complementary. The ML model was trained on the TCGA data, and then applied to the PDX datasets. The two most popular ML models used to train based on TCGA are Support Vector Machines (SVMs) and Random Forest (RF) (tree based)[5]; Certis compared SVM and RF with logistic regression in Table 1 (top right) and found a slight benefit from RF versus the other two for Certis PDX.

Data processing: Certis used TCGA data as the training dataset, and two different test datasets, both of which were PDX: the BarneyOI Cancer Model Database™ and the NCI Patient Derived Models Repository (PDMR)[6]. Certis limited the TCGA training data to the cancer types represented in the PDX datasets. Gene expression transcript per million (TPM) values were quantile normalized with the training data before dimension reduction derived from training data were applied to the test data. Finally, the ML model which was trained on TCGA data was applied to the dimension reduced PDX test data.

Gene Expression Analysis — RNA Sequencing (RNA-Seq): Gene expression was measured experimentally by bulk poly(A)-selected RNA-Seq. Mouse contamination was removed (Xenome v1.0) and further processed for quality using fastp and FastQC. STAR was used to map stranded paired-end reads to the Human GRCh38 (hg38) genome, and gene expression was quantified using RSEM.

ACCESS BARNEYOI™ CANCER MODEL DATABASE
certisoncology.com/tumor-bank

RESULTS



PDX Data	ML Model	Accuracy	Precision	Recall
Certis	Random Forest	0.87	0.91	0.87
Certis	Support Vector Machine	0.86	0.90	0.86
Certis	Logistic Regression	0.86	0.90	0.86
PDMR	Random Forest	0.82	0.89	0.82
PDMR	Support Vector Machine	0.85	0.92	0.85
PDMR	Logistic Regression	0.85	0.93	0.85

Table 1. Comparison of ML models. Three different ML models were trained on the results of the dimension reduction algorithms (PCA and UMAP). The predictions for each sample were then combined and each class was averaged (weighted) to obtain the global statistics shown. Certis found that after dimension reduction, the ML models performed very similarly with Random Forest and had a slight benefit for Certis PDX models.

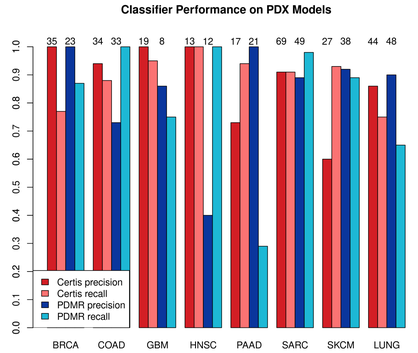


Figure 2. Performance of ML classifier on PDX (test) datasets. Recall (true positive rate) and precision (Positive Predictive Value, or 1 – False Discovery Rate) are shown for both Certis and PDMR PDX datasets. In a multiclass classification ML setting, false positives (FP) and false negatives (FN) are only defined per class, rather than globally. For this reason, the results for each class may appear lower than the overall performance for Certis of 91% precision and 87% recall, or 87% accuracy.

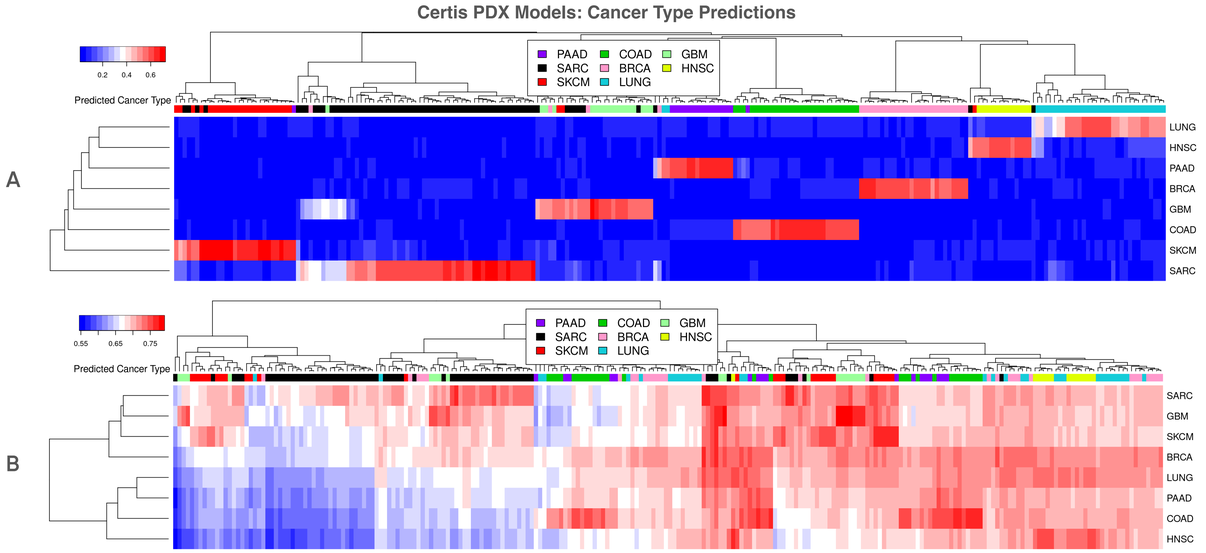


Figure 1. Clustering (heatmap) analysis of cancer type predictions for Certis PDX models:
A. Using a probability estimator based on the random forest ML model, probabilities for each cancer type (rows) for each PDX model (columns) are shown in the heatmap. A bar above the plot shows the diagnosed (annotated) cancer type. The probabilities obtained from training on UMAP features were averaged with the probabilities based on PCA features for display purposes.
B. Correlation analysis: To put the previous plot in context, Certis plotted the gene level TPM spearman correlation between each model and the median TPM of TCGA data for each cancer type (TPM > 0.1, roughly 20K genes). Accuracy and separation between classes were approximately 20% worse compared to the ML models above.

CONCLUSIONS

The high degree of concordance between diagnosed and predicted cancer types provides confidence that PDX models accurately recapitulate patient tumors. Remaining discrepancies are likely due to heterogeneity of training and test data, including inaccurate diagnoses, metastases, treatment effects, etc., and are expected in such a complex phenotype as cancer. To see detailed next-generation sequencing (NGS) data on all cancer models in BarneyOI, register to access our searchable database.

CITATIONS & ACKNOWLEDGEMENTS

[1] Certis Oncology Solutions, San Diego, CA.
[2]Chin DH et. al. *Pharmaceuticals* (2023).
[3]Clayton EA et al. *BMC Bioinformatics* (2020).
[4]Yang Y et al. *Cell Reports* (2021).
[5]Liñares-Blanco J et al. *PeerJ Comput Sci.* (2021).
[6]The NCI PDMR, NCI-Frederick, *Frederick National Laboratory for Cancer Research*, Frederick, MD.