Multi-cancer Brain Metastasis Risk Score Development and Validation using 220,246 Whole Transcriptomes and Machine Learning

[1]Jim Abraham, [2]Carey K. Anders, [3]Adam Brufsky, [4]Michael J. Glantz, [5]Priscilla Kaliopi Brastianos, [6]Luke Roy George Pike, [7]Amy B. Heimberger, [1]George W. Sledge Jr., [1]Matthew Oberley, [1]David Spetzler

[1]Caris Life Caris Life Sciences, Phoenix, AZ; [2]Duke Cancer Institute, Durham, NC; [3]Magee-Womens Hospital of UPMC/ UPMC Hillman Cancer Center, Pittsburgh, PA; [4]Penn State Milton S. Hershey Medical Center, Hershey, PA; [5]Massachusetts General Hospital, Boston, MA; [6]Memorial Sloan Kettering Cancer Center, New York, NY; [7]Northwestern University, Chicago, IL

CARIS® RESEARCH INSTITUTE

jabraham@carisls.com | Abstract ID: 2039

Introduction:

Application of AI to large molecular data sets covering transcribed genes is in its infancy. To date, there have not been large enough data sets to take advantage of the power of AI technology to predict disease progression for patients with cancer. Here we show that application of ML/AI methodologies to a large collection of molecular and clinical data, generates insight into disease progression. While meaningful predictions can be made on cohorts of 100,000 patients, it is clear that more data will enable even more accurate predictions. Generation of WTS data on larger patient cohorts is essential to maximize precision medicine and advance the science and medicine of cancer care. These results can have a direct impact on current patients as well as provide insight into future drug develop efforts.

Development

AI Feature Generation Layer

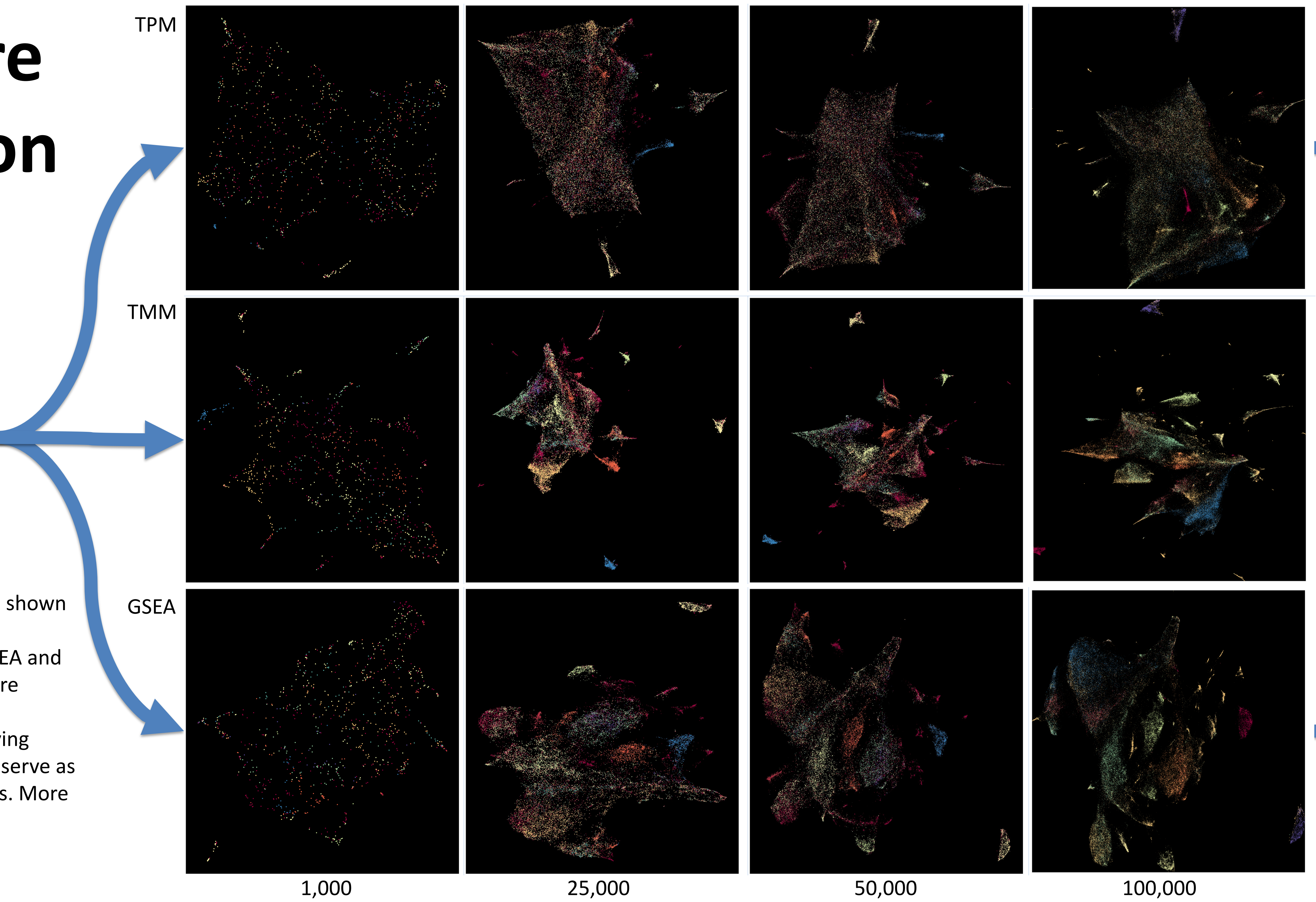

Figure 1: Embedding of WTS data is shown being applied to patients using a combination of TPM, TMM, and GSEA and delivered to three layers of AI feature generation algorithms to reduce dimensionality and identify underlying clusters where the coordinates can serve as inputs into secondary AI/ML models. More data results in more refined cluster coordinates.

AI Model Generation Layer

Figure 2: Features from the primary AI layers are used as inputs into a multi-layer deliberation analytic AI platform which learns patients that are more or less likely to develop a brain metastasis. There were 7 populations identified based on risk probabilities.

Validation

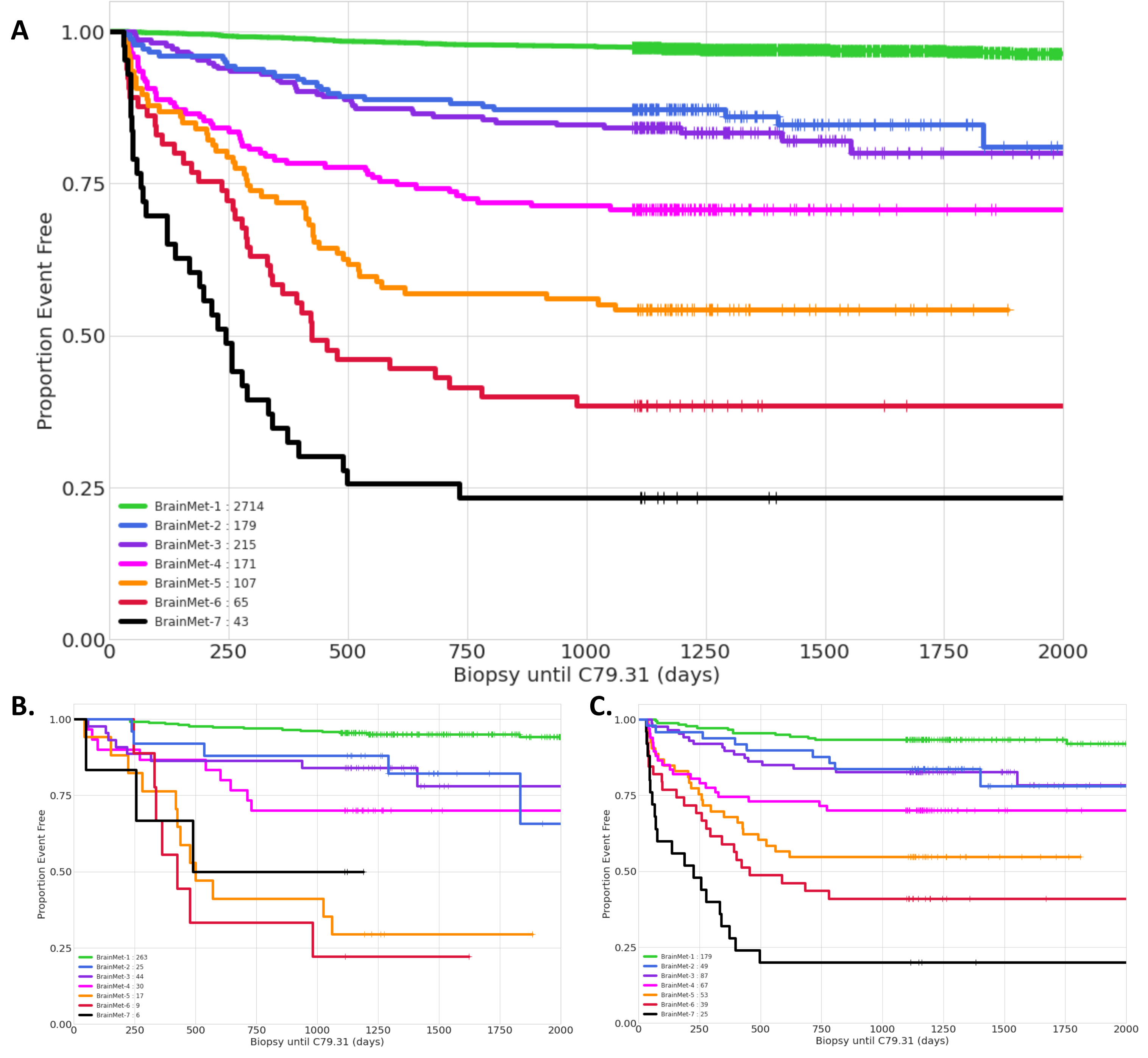

Figure 3: Predictions on an independent set of patients show the relative likelihood of developing a brain met with A) all B) breast, and C) NSCLC cancers. All samples are from the primary tumor site. In A) the median time to brain met development for groups 1-5 was unevaluable while groups 6-7 were 424 and 244, respectively.

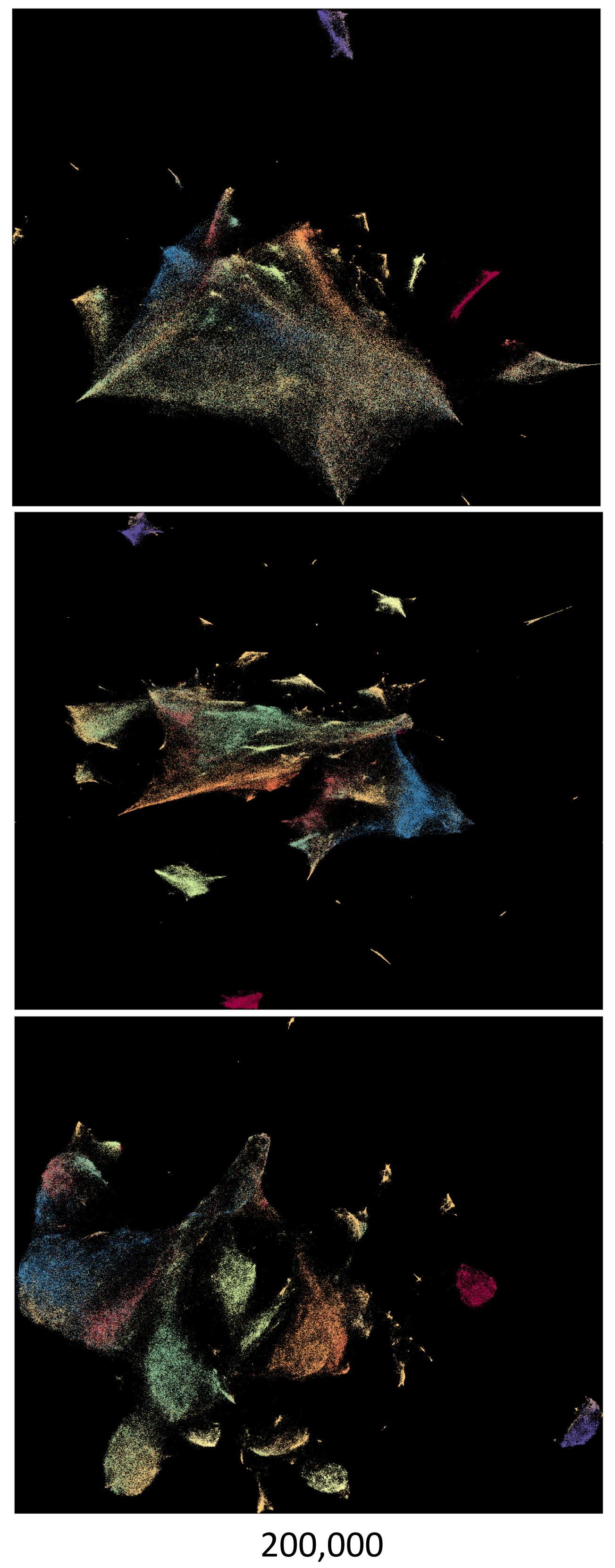

Figure 4: Embeddings generated using all the data demonstrate that the coordinates used to validate the brain met predictor could be improved upon as there is further refinement in clusters.

Background

Brain metastases occur in multiple cancer types with higher prevalence in lung, breast, melanoma, and GI cancers[2]. The prognoses of patients who develop brain metastases are very poor and identification of brain metastasis risk could be useful for prognostication, monitoring, and therapy selection.

Methods

Data from the whole transcriptome of 220,246 tumor profiles were analyzed and multiple machine learning models were trained on various molecular subtypes. The dataset was split 50% for training and the other 50% for testing, UMAP[1] was employed for dimensionality reduction and the patterns learned across the entirety of the training dataset irrespective of brain metastasis were leveraged on the testing data set. Patients with brain metastasis were identified using the presence of ICD-10 code C79.31 (Secondary malignant neoplasm of the brain). As the absence of C79.31 could be due to the event not happening yet, patients without brain metastasis were stratified into groups based on 3, 4 or greater than 5 years without a C79.31 ICD-10 code. The brain metastasis risk score was defined by empirical evaluation of the positive predictive value in 7 groups of risk probabilities. The validation set contained 1,217 patients with brain metastasis and 4,631 without an observed brain metastasis within 3 years.

Results

In the validation set, the prevalence of brain metastases within the risk scores across all cancer types ranged from 4% with the lowest risk score to 94% in the highest with 71% of cases receiving the lowest 2 risk scores, 15% the 2 intermediate risk scores, and 14% the 3 highest risk scores. For breast, lung and colon cancers, the prevalence of brain metastasis ranged from 4-10% in patients with the lowest risk scores to 92-100% in patients with the highest however the distribution of cases with each risk score was markedly different across cancer type. Breast cancer had 62% of cases receiving the lowest 2 risk scores versus 27% in lung, and 92% in colon. Breast cancer had 18% of cases receiving the 3 highest risk scores while lung had 42% and colon only 2% of cases with those 3 highest scores.

Conclusions

Whole transcriptome data can be leveraged by a machine learning platform that employs dimensionality reduction techniques along with transfer learning to predict the risk of brain metastasis. This tool can be used to augment the clinical picture of cancer patients an unmet clinical opportunity to inform prognosis, monitoring, and therapeutic selection.

References

1. McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018
2. Abdulkarim et al., JCO. V 29. n11, 2011
3. Cosgrove et al., Nature Communications. 13. 514. 2022
4. Khan et al., Nature Medicine. 7, 673-679, 2001
5. Lopez-Garcia et al., PLOS. 2020