Machine Learning Algorithm Analysis using 55,780 Cases from a Commercial 592-gene NGS Panel to Accurately Predict Tumor Type for Carcinoma of Unknown Primary (CUP)

[1]Jim Abraham, [2]Amy B. Heimberger, [3]John Marshall, [1]Joanne Xiu, [1]Anthony Helmstetter, [1]Daniel Magee, [1]Adam Morgan, [1]Curtis Johnston, [1]Zoran Gatalica, [1]Wolfgang Michael Korn, [1]David Spetzler

[1]Caris Life Sciences, Phoenix, AZ, [2]The University of Texas MD Anderson Cancer Center, Department of Neurosurgery, Houston, TX, [3]Georgetown Lombardi Cancer Center, Washington, DC

Abstract

Background: The diagnosis of a malignancy is typically informed by clinical presentation and tumor tissue features including cell morphology, immunohistochemistry, cytogenetics, and molecular markers. However, in approximately 5-10% of cancers[1,2], ambiguity is high enough that no tissue of origin can be determined and the specimen is labeled as a Cancer of Occult/Unknown Primary (CUP). Lack of reliable classification of a tumor poses a significant treatment dilemma for the oncologist leading to inappropriate and/or delayed treatment. Gene expression profiling has been used to try to identify the tumor type for CUP patients, but suffers from a number of inherent limitations. Specifically, tumor percentage, variation in expression, and the dynamic nature of RNA all contribute to suboptimal performance. For example, one commercial RNA-based assay has sensitivity of 83% in a test set of 187 tumors and confirmed results on only 78% of a separate 300 sample validation set[3].

Methods: 55,780 tumor patients with NGS data were used to construct a multiple parameter tumor type specific classification system using an advanced machine learning approach.

Tumor Types with Specific DNA Aberrations

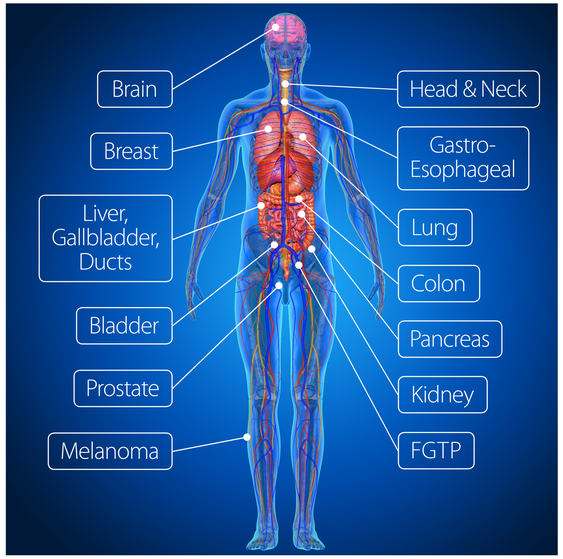

Figure 1 – Tumor types that the algorithm is able to differentiate.

DNA Can Identify Tumor Type

Sensitivity	Specificity	PPV	NPV	Accuracy	Call Rate
90.0%	98.0%	90.0%	98.6%	97.3%	97.5%

Table 1 – Performance metrics of assay on an independent test set of 15,473 cases.

Robust to Metastasis and Tumor Percentage

	Sensitivity	Specificity	PPV	NPV	Accuracy	Call Rate
Primary	90.9%	98.0%	91.1%	98.9%	97.6%	97.3%
Metastatic	89.0%	97.9%	89.3%	98.2%	96.9%	97.6%
20-50% Tumor	90.3%	98.2%	90.6%	98.5%	97.5%	97.1%
>50% Tumor	90.3%	98.2%	90.6%	98.5%	97.5%	97.1%

Table 2 – Performance metrics on subsets of the test data from a primary site (N = 8,437), metastatic site (6,690), and samples with low (9,492) and high tumor percentages (5,945).

Performance Holds Across Multiple Tumor Types

Tumor Type	Train N	Test N	Sensitivity	Specificity	PPV	NPV	Accuracy	Call Rate
Head, Face, Neck	299	144	45.4%	100.0%	96.4%	99.6%	99.6%	82.6%
Melanoma	976	402	85.0%	99.9%	94.3%	99.6%	99.5%	96.3%
FGTP	8,872	4,115	93.4%	98.3%	95.4%	97.6%	97.0%	98.8%
Prostate	785	477	96.1%	99.8%	94.7%	99.9%	99.7%	96.6%
Brain	1,554	479	93.3%	99.8%	93.5%	99.8%	99.6%	96.0%
Colon	5,805	2,532	94.5%	98.5%	92.9%	98.9%	97.9%	98.9%
Kidney	426	178	84.1%	99.9%	91.7%	99.8%	99.8%	88.2%
Bladder	447	304	60.6%	99.9%	89.4%	99.3%	99.1%	91.8%
Breast	3,324	1,386	90.9%	98.7%	87.9%	99.1%	98.0%	98.3%
Lung	7,744	3,540	96.0%	95.4%	86.3%	98.7%	95.5%	98.2%
Pancreas	1,637	708	83.7%	99.3%	84.6%	99.2%	98.5%	98.3%
Gastroesophageal	1,521	743	72.0%	99.3%	82.6%	98.6%	98.0%	93.8%
Liver, Gallbladder, Ducts	734	364	57.7%	99.7%	82.2%	99.0%	98.8%	92.6%

Table 3 – Performance metrics and cohort sizes of subsets of the independent test dataset where the primary tumor site is known.

Algorithmic Classification of CUP Cases

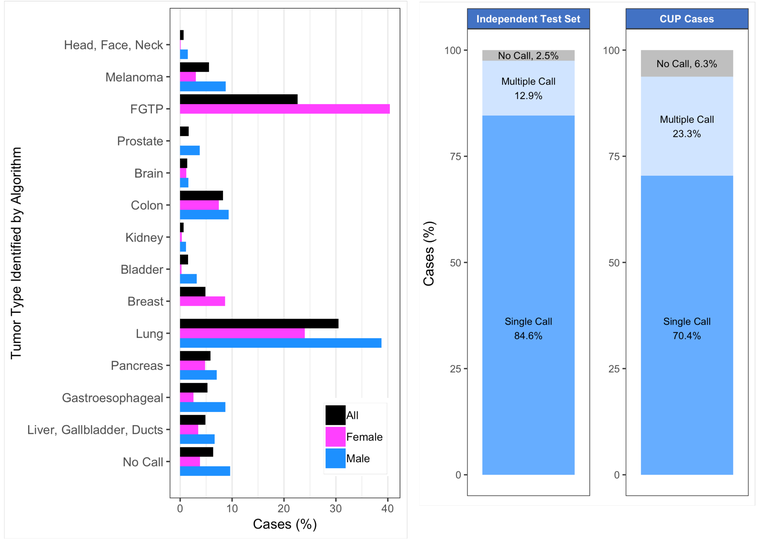

Figure 2 – Percentages of tumor types identified in CUP cases (N = 5,955) by the algorithm (left panel) and the distribution of required calls in order to identify a tumor type with a sufficiently high probability (right panel).

Conclusions

- Final performance of DNA-based tumor type identification on an independent test of 15,000+ patient samples is superior to current standards using gene expression based methods
- Unbiased training machine learning techniques applied to more than 45,000 enabled detection of tumor types independent of sampling location or tumor percentage
- Tumor type predictors can render a histologic diagnosis to CUP cases that can inform treatment and potentially improve outcomes
- Cancer of unknown primary remains a substantial problem for both clinicians and patients, diagnosis can be aided with the algorithms presented here.
- Returning both diagnostic and therapeutic information that optimize patients treatment strategy from a single test is a substantial improvement over the current standard of multiple tests that require more tissue

References

1. https://www.mdanderson.org/cancer-types/cancer-of-unknown-primary.html
2. https://www.cancer.gov/types/unknown-primary/hp/unknown-primary-treatment-pdq#_1
3. Erlander MG, et al. Performance and clinical evaluation of the 92-gene real-time PCR assay for tumor classification. J Mol Diagn. 2011 Sep;13(5):493-503. doi: 10.1016/j.jmoldx.2011.04.004. Epub 2011 Jun 25.

Distribution of Calls Per Tumor Type Indicate Unique Molecular Subtypes

Tumor Type Identified by Algorithm \ True Tumor Type	Head, Face, Neck	FGTP	Prostate	Melanoma	Brain	Colon	Kidney	Bladder	Breast	Lung	Pancreas	Gastro-esophageal	Liver, Gallbladder, Ducts
Head, Face, Neck	0.96	0	0	0	0	0	0	0.04	0	0	0	0	0
Melanoma	0.01	0.96	0.01	0	0	0.01	0	0	0	0	0	0	0
FGTP	0	0	0.95	0	0	0.01	0	0	0.01	0.01	0	0	0
Prostate	0.01	0	0	0.95	0	0	0	0.01	0	0.01	0	0	0
Brain	0	0	0.03	0	0.93	0	0	0	0.01	0	0	0.01	0
Colon	0	0	0.01	0	0	0.93	0	0	0	0.01	0.01	0.02	0.01
Kidney	0	0	0.01	0	0	0.01	0.92	0	0	0.03	0.01	0.01	0.01
Bladder	0.01	0.01	0.02	0	0	0	0.01	0.89	0.01	0.03	0.01	0.01	0.02
Breast	0	0	0.07	0	0	0	0	0.01	0.88	0.01	0	0	0.01
Lung	0.01	0.01	0.02	0	0	0.01	0	0.01	0.01	0.86	0.02	0.02	0.01
Pancreas	0	0	0.02	0	0	0.03	0	0	0	0.03	0.85	0.02	0.05
Gastroesophageal	0.01	0	0.02	0	0	0.05	0	0.01	0.01	0.03	0.01	0.83	0.02
Liver, Gallbladder, Ducts	0	0	0.02	0	0	0.01	0.01	0.01	0.02	0.03	0.04	0.03	0.82

Table 4 – Proportion of cases identified by the algorithm for each tumor type.