The American Journal of Pathology, Vol. ■, No. ■, ■ 2022

ELSEVIER

The American Journal of PATHOLOGY

ajp.amjpathol.org

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

Hong Zhang,* Muhammad A. Qureshi,† Mohsin Wahid,† Ahmad Charifa,* Amir Ehsan,‡ Andrew IP,§ Ivan De Dios,* Wanlong Ma,* Ipsa Sharma,* James McCloskey,§ Michele Donato,§ David Siegel,§ Martin Gutierrez,§ Andrew Pecora,§ Andre Goy,§¶ and Maher Albitar*¶

From the Genomic Testing Cooperative, Irvine, California; the Dow University of Health Sciences Karachi,† Karachi City, Pakistan; the CorePath Laboratories,‡ San Antonio, Texas; the John Theurer Cancer Center,§ Hackensack Meridian Health, Hackensack, New Jersey; and the Hackensack Meridian School of Medicine,¶ Nutley, New Jersey*

Accepted for publication September 20, 2022.

Address correspondence to Maher Albitar, Genomic Testing Cooperative, 175 Technology Dr. #100, Irvine, CA 92186.
E-mail: malbitar@genomictestingcooperative.com.

Diagnosis and classification of tumors is increasingly dependent on biomarkers. RNA expression profiling using next-generation sequencing provides reliable and reproducible information on the biology of cancer. This study investigated targeted transcriptome and artificial intelligence for differential diagnosis of hematologic and solid tumors. RNA samples from hematologic neoplasms ($N = 2606$), solid tumors ($N = 2038$), normal bone marrow ($N = 782$), and lymph node control ($N = 24$) were sequenced using next-generation sequencing using a targeted 1408-gene panel. There were 20 subtypes of hematologic neoplasms and 24 subtypes of solid tumors. Machine learning was used for diagnosis between two classes. Geometric mean naïve bayesian classifier was used for differential diagnosis across 45 diagnostic entities with assigned rankings. Machine learning showed high accuracy in distinguishing between two diagnoses, with area under the curve varying between 1 and 0.841. Geometric mean naïve bayesian algorithm was trained using 3045 samples and tested on 1415 samples, and showed correct first-choice diagnosis in 100%, 88%, 85%, 82%, 88%, 72%, and 72% of acute lymphoblastic leukemia, acute myeloid leukemia, diffuse large B-cell lymphoma, colorectal cancer, lung cancer, chronic lymphocytic leukemia, and follicular lymphoma cases, respectively. We conclude that targeted transcriptome combined with artificial intelligence are highly useful for diagnosis and classification of various cancers. Mutation profiles and clinical information can improve these algorithms and minimize errors in diagnoses. *(Am J Pathol 2022, ■: 1–9; https://doi.org/10.1016/j.ajpath.2022.09.006)*

Diagnosis and classification of tumors are increasingly dependent on biological and molecular biomarkers. The management and therapy of cancer vary significantly, depending on the proper classification of cancer. Relying on the expertise of a pathologist and the morphologic features of the tumor alone lead to significant discrepancies in diagnosis because of the subjective nature. More important, the possibility of an incorrect diagnosis is relatively high. Numerous studies have shown that errors in the diagnosis and classification of cancers continue to be a significant issue in current clinical practice.[1–6] Recent advances in utilizing machine learning to determine the morphology and immunohistochemistry of tumors are promising for improving cancer diagnosis and classification and reducing variability.[7–9]

Supported by the Genomic Testing Cooperative.

Disclosures: M.A., H.Z., A.C., I.D.D., and W.M. work and own stocks in a diagnostic company that offers RNA sequencing using artificial intelligence. J.M. served on a speaker's bureau for Amgen, Bristol Myers Squibb, Incyte, Jazz Pharmaceuticals, Stemline, and Takeda; and has served as a consultant for AbbVie, CTI BioPharma, and Novartis. A.G. has consulting/advisory board/honoraria from AstraZeneca, SecuraBio, and TG Therapeutics, not relevant to this work. M.W., A.I., M.D., D.S., M.G. and A.P. have no relevant conflict of interest.

Copyright © 2022 American Society for Investigative Pathology. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).
https://doi.org/10.1016/j.ajpath.2022.09.006

UNCORRECTED PROOF

RNA profiling of cancer cells has been highly useful for providing information on the tumor, microenvironment, and immune response.[10,11] Using next-generation sequencing to analyze RNA enables profiling a reliable clinical tool and approach for the discovery of biomarkers, characterizing the biology of tumors, and predicting the efficacy of various therapeutic approaches.[10,11] RNA sequencing and quantification using next-generation sequencing is more reliable and reproducible than old technologies, such as microarrays or PCR-based RNA quantification.[12–14] Targeted RNA sequencing of various tissue samples allows us to focus on relevant oncogenic markers and to sequence at a deeper level for better quantification of low-level expressor genes that might be major regulators of the complex biology of cells. This study explored the potential of using targeted transcriptomes and artificial intelligence for the differential diagnosis and classification of hematologic and solid tumors.

Materials and Methods

Patients and Samples

A total of 5450 fresh bone marrow (BM) and formalin-fixed, paraffin-embedded (FFPE) cancer samples were obtained for this study (Supplemental Table S1). The samples were collected consecutively without any selection during routine clinical molecular profiling using next-generation sequencing of DNA and RNA between November 2018 and November 2021. The tumor fractions varied between 30% and 80%, and the samples reflected a real-time occurrence. Diagnoses of the samples are listed in Table 1. The samples for various types of leukemia, myelodysplasia, and normal tissues were collected from fresh BM. On the other hand, lymphoma cases and solid tumors were based on FFPE samples. Tumor diagnosis was confirmed using morphologic analysis, flow cytometry, immunohistochemistry, and molecular profiling of the DNA and RNA. The DNA and RNA were extracted from the BM samples using an automated Maxwell System platform (Promega, Madison, WI). Agencourt FormaPure Total 96-Prep Kit was used to extract both the DNA and RNA from FFPE samples using an automated KingFisher Flex, following the manufacturer's recommendations. Agencourt FormaPure Kit had provision for a split protocol to extract both the DNA and RNA from the same FFPE lysate. Blood and BM samples were collected in EDTA. The DNA and RNA were extracted from fresh samples within 72 hours of collection. The study protocol was approved by Institutional Review Board by Western Copernicus Group (New England Institutional Review Board, Aspire Institutional Review Board, and Midlands Institutional Review Board; number 1-1476184-1). Consent was waived because of incidental collection and lack of risk. This study was conducted in accordance with the principles of the Declaration of Helsinki and its later amendments.

Table 1 List of Neoplasms and Samples

Disease	*N*
Aplastic anemia	12
Acute lymphoblastic leukemia	89
Acute myeloid leukemia	352
Brain tumors	44
Breast cancer	137
Burkitt lymphoma	10
Carcinoma (not otherwise specified)	32
Clear cell renal cell carcinoma	8
Cholangiocarcinoma	9
Chronic lymphocytic leukemia	167
Chronic myeloid leukemia	46
Chronic myelomonocytic leukemia	97
Colorectal carcinoma	308
Diffuse large B-cell lymphoma	746
Endometrial cancer	113
Esophageal carcinoma	34
Follicular lymphoma	145
Gallbladder carcinoma	4
Gastric carcinoma	10
Gastrointestinal stromal tumor	11
Hairy cell leukemia	5
Head and neck tumor	4
Hodgkin lymphoma	65
Lung cancer	794
Lymphoma (not otherwise classified)	3
Mantle cell lymphoma	93
Marginal zone lymphoma	76
Myelodysplastic syndrome	316
Melanoma	21
Multiple myeloma	113
Myeloproliferative neoplasms	88
Neuroendocrine tumor	5
Normal bone marrow, fresh	782
Normal lymph node	24
Ovarian cancer	126
Pancreatic cancer	96
Prostate cancer	36
Sarcoma	137
Squamous cell carcinoma of skin	15
T-cell acute lymphoblastic leukemia	7
T-cell lymphoma	145
Thyroid cancer	24
Upper gastrointestinal cancer	23
Urothelial cancer	38
Vulva cancer	9
Waldenstrom macroglobulinemia	31
Total	5450

RNA Library Construction and Sequencing

The samples were selectively enriched for 1408 cancer-associated genes using the reagents provided in the Illumina TruSight RNA pan-cancer panel (Illumina, San Diego, CA) (Supplemental Table S1). cDNA was generated from the cleaved RNA fragments using random primers during the first- and second-strand synthesis. Sequencing adapters were

UNCORRECTED PROOF

ARTICLE IN PRESS

ligated into the resulting double-stranded cDNA fragments. The coding regions of the expressed genes were captured from this library using sequence-specific probes to generate the final library. Sequencing was performed using the Illumina NextSeq 550 system platform. Ten million reads per sample were performed in a single run, and the read length was 2 × 150 bp. An expression profile was generated from the sequencing coverage profile of each sample using Cufflinks. Expression levels were measured as fragments per kilobase of transcripts per million.

Using Machine Learning Algorithm for Classification of Two Diagnostic Classes

The RNA expression data were used in the machine learning algorithm to distinguish between any two diagnostic classes. Recognizing that high dimensionality of the problem would make it vulnerable to overfitting with many artificial intelligence techniques, we addressed this problem by applying a modified version of naïve Bayes (geometric mean naïve Bayes). The conditional independence assumption of the naïve Bayes is almost never strictly satisfied in practical applications. However, it is still a useful tool, especially in situations like this problem. With a high dimension and a limited sample size, to estimate the correlations between genes would be counterproductive. Naïve Bayes approach has a small number of parameters and, hence, a lower capacity as a learning system, which will help address the overfitting problem according to statistical learning theory. We developed the geometric mean naïve Bayes method to address the numeric underflow issue of standard naïve Bayes when applied to a high-dimensional problem. When the likelihood is the product of thousands of conditional probabilities, underflow is unavoidable, even with the proportional scaling. In geometric mean naïve Bayes, we apply the geometric mean to the conditional probabilities. The method is documented in a separate article.[15] We proved that the geometric mean is essentially the only operation that will preserve the conditional independence of naïve Bayes and will not cause underflow. The gene selection/filtering method is used to eliminate irrelevant genes and improve the training. Two statistical criteria on individual gene were applied to perform the filtering. The two measures are applied to the individual gene for the sole purpose of eliminating irrelevant genes. Consequently, they are not used to measure performance and confidence of the final classifier. The two measures provide indications on how relevant a gene is to distinguish the classes. Although they are used for the same purpose, the two measures do not give the same ranking on the genes. To explain the rationale of defining such measures, we used the terms of performance measure and stability measure. In selecting genes that distinguished between the two classes, we used the standard naïve bayesian classifier on each gene with k-fold cross-validation.

$$a = \frac{1}{m}\sum_{i=1}^{m}\frac{t_i}{n_i} \tag{1}$$

where m is the number of classes, n_i is the number of cases in the class, and t_i is the number of correctly classified cases in class i estimated using the k-fold cross-validation.

When the number of classes was $m = 2$, the measures were the average of the sensitivity and specificity. In general, this was the average of the accuracies of the individual classes. Overall accuracy is not appropriate for gene selection, because it can be misleading in data sets with unbalanced classes (eg, in a data set with 80 negative and 20 positive cases, a trivial classification of all negatives yields 80% accuracy). The coefficient $1/m$ was usually ignored in our study, because m was a constant and did not affect the ranking. The k-fold cross-validation was usually implemented with $k = n$ (ie, leave one out). Although the leave-one-out method was computationally more extensive, the efficiency of the naïve Bayes algorithm still made the selection process reasonably fast. This accuracy value provided a direct measure of the genes used for classifying groups; however, this did not provide confidence information. For the confidence measure for gene selection, we relied on the P value of a gene to differentiate the classes. Analysis of variance was applied to compute the P value for a gene to discriminate between groups.

$$F = \frac{MSB}{MSW} \tag{2}$$

where MSB was the mean sum of squares between groups, MSW was the mean sum of squares within groups, and F was the analysis of variance coefficient following the F distribution. The P value was obtained from the F value. This confidence value provided the measure of the stability and robustness of the gene in the classifying groups. It did not provide concrete classification accuracy but contributed the overall confidence in the differences of the class means. Both criteria provided quantitative measures of the relevance of a gene for classification; however, these two relevance measures did not always produce the same ranking. Applying both measures would produce effective and stable gene selection methods for machine learning–based classification systems.

After selecting individual genes, we used a naïve bayesian classifier to distinguish between diagnostic classes using multiple selected genes with both confidence and P values. However, because the naïve bayesian classifier has severe numerical underflow problems when the dimensions of data were high, we developed the geometric mean naïve bayesian (GMNB) classifier that eliminated the underflow problem by applying a multiplicative positive increasing function to the likelihood. In particular,

$$P(x_1, x_2, \ldots, x_d|C_j) = \sqrt[d]{[P(x_1|C_j)P(x_2|C_j)\ldots P(x_d|C_j)]} \tag{3}$$

UNCORRECTED PROOF

FLA 5.6.0 DTD ■ AJPA3794_proof ■ 19 October 2022 ■ 10:22 am ■ EO: AJPA-D-22-

This formula represented the geometric mean of conditional probabilities. We proved that the GMNB method resolved the underflow problem for high-dimensional data by showing that the expected value of such a likelihood approached $1/e$ when dimension $d \rightarrow \infty$. We also proved that such a function is unique up to a constant multiple of exponent.[14]

To reduce the effects of noise and avoid overfitting when selecting these genes, we employed leave-one-out cross-validation to obtain a robust performance measure. For an individual gene, a GMNB was constructed on the training subset and examined on the testing subset. The complement of the cross-validation error rate was used as the discriminant measure for bins.

$$d = \sum_{c=1}^{k} 1 - \frac{error_c}{n_c} \quad (4)$$

Instead of the overall error rate, the value d takes the sum of the error rates of individual classes. This definition avoided bias when the sample sizes were not balanced for different classes. The genes were ranked by d, with higher values corresponding to better-performing genes for classification. To address stability issues, we used t-test to measure the significance of a bin separating the two classes. By setting a P-value threshold, insignificant bins can be filtered. The selected genes were used to distinguish between the two classes using a k-fold cross-validation procedure (with k = 12). A naïve bayesian classifier was constructed for the training of k-1 subsets and evaluated on the other testing subset. The training and testing subsets were then rotated, and the average of the classification errors was used to measure the relevance of the gene. The classification system was trained using a selected subset of the most relevant genes. The processes of gene selection and class selection were applied recurrently to obtain an optimal classification system, and a subset of genes relevant to distinguishing between the two classes was defined and isolated.

Using GMNB Classifier for Ranking Diagnostic Classes

GMNB classifier described above was also used in classifying each sample against multiple diagnostic classes. The GMNB method resolved the underflow problem of high-dimensional data.

Results

High Accuracy in the Differential Diagnosis between Two Diagnostic Classes

Initially, we evaluated the ability of machine learning to distinguish between the two disease classes. Using a machine learning algorithm, we first selected the proper genes to distinguish between the two classes by using the best classifier biomarkers based on the P value for predicting a specific diagnosis. This approach showed high sensitivity and specificity for distinguishing between the two diagnoses (Table 2). As shown in Figure 1 and Table 2, area under the curve for most classifications was >0.90. Distinguishing between normal and myelodysplastic syndrome (MDS) was relatively less reliable because of the significant overlap between the two entities. The algorithm used the expression of 400 genes to achieve a sensitivity of 78.1% and a specificity of 75.3%. However, the presence or absence of mutations can easily distinguish between these entities. Similarly, distinguishing between MDS and myeloproliferative neoplasms (MPNs) was relatively less robust because of the known overlap between the two entities. Using the expression of 500 genes, the algorithm achieved a sensitivity of 90.9% and a specificity of 70.8%. This is without using mutation profiles. As expected, using mutation profiling can reliably distinguish between these entities. Furthermore, clinical cases with features of both MDS and MPN are well documented, and the relatively poor distinction between these two entities might be due to the presence of such cases between the samples used in this study. Distinguishing between chronic lymphocytic leukemia, mantle cell lymphoma, and marginal zone lymphoma was remarkably reliable (Table 2). The expression of mere 10 genes was adequate to distinguish between chronic lymphocytic leukemia and mantle cell lymphoma, with a sensitivity of 94.6% and a specificity of 95.2%. Distinguishing between chronic lymphocytic leukemia and marginal zone lymphoma required the expression profile of 25 genes to achieve a sensitivity of 98.7% and a specificity of 91%, due to the overlap between the two entities. Furthermore, distinguishing between Hodgkin lymphoma and normal lymph node, or T-cell lymphoma, was highly reliable using the expression of 100 and 500 genes, respectively (Table 2). Similarly, distinguishing between various solid tumors was also highly reliable. Distinguishing between various sarcoma cases and gastrointestinal stromal tumors was highly reliable using the expression of 100 genes.

Differential Diagnosis between 47 Different Diagnostic Classes with Ranking

Distinguishing between multiple classes was significantly more complex, because the specific biomarkers that are suitable for distinguishing between two diagnostic classes may not be relevant for determining the differences between these two classes and the rest of the diagnostic classes.

The most difficult initial step was the selection of biomarkers for distinguishing between one class and the rest of the 47 diagnostic classes. To overcome this problem, we used all 1408 biomarkers without selection to provide a score that could be ranked to predict a specific diagnostic class. We used a machine learning approach that is based on a generalized naïve bayesian classifier to train the system to distinguish between 47 different diagnostic classes. We first used 3045 cases for training; then, we used 1415 cases for

UNCORRECTED PROOF

Table 2 Transcriptome and Differential Diagnosis between Two Diagnostic Classes

Two classes	AUC (95% CI)	Sensitivity, %	Specificity, %	Genes, *N*	AUC − 1 (95% CI)
Normal versus AML	0.9764 (0.954–0.974)	90.9	93.2	100	0.945 (0.933–0.957)
Normal versus ALL	0.981 (0.973–0.989)	95.1	95.5	200	0.977 (0.968–0.985)
Normal versus CLL	0.997 (0.994–0.999)	96.4	98.8	100	0.980 (0.973–0.988)
Normal versus mantle	0.992 (0.987–0.997)	95.1	97.8	100	0.969 (0.959–0.980)
Normal versus MDS	0.831 (0.801–0.861)	78.1	75.3	400	0.826 (0.796–0.856)
Normal versus MPN	0.923 (0.884–0.962)	90.9	82.3	400	0.903 (0.860–0.946)
MDS versus MPN	0.884 (0.837–0.931)	90.9	70.8	500	0.806 (0.748–0.864)
AML versus MDS	0.880 (0.854–0.906)	86.1	70.2	400	0.864 (0.837–0.892)
CLL versus mantle	0.986 (0.968–1.000)	94.6	95.2	10	0.986 (0.968–1.00)
Marginal versus CLL	0.984 (0.964–1.00)	98.7	91	25	0.864 (0.809–0.920)
Marginal versus follicular	0.946 (0.917–0.974)	91	93.4	550	0.942 (0.912–0.971)
Hodgkin versus normal LN	0.990 (0.972–1.00)	95.4	100	100	1.00 (1.00–1.00)
Hodgkin versus T-cell lymphoma	0.963 (0.930–0.996)	92.3	91	500	0.902 (0.850–0.954)
Hodgkin versus DLBCL	0.975 (0.948–1.00)	96.9	95.3	500	0.965 (0.934–0.997)
DLBCL versus follicular	0.986 (0.972–0.999)	95.9	93.1	600	0.975 (0.957–0.993)
DLBCL versus T-cell lymphoma	0.967 (−0.946 to 0.988)	91.7	89.8	600	0.942 (0.915–0.969)
Lung versus colorectal	0.982 (0.975–0.989)	97.2	94.5	900	0.977 (0.969–0.985)
Lung versus breast	0.988 (0.982–0.994)	98	92.7	700	0.988 (0.982–0.994)
Breast versus ovarian	0.994 (0.984–1.00)	100	94.2	700	0.989 (0.976–1.00)
Ovarian versus endometrial	0.959 (0.933–0.984)	92.9	91.2	600	0.853 (0.803–0.902)
Breast versus colorectal	0.997 (0.991–1.00)	97.8	98.7	800	0.987 (0.973–1.00)
Pancreas versus colorectal	0.989 (0.980–0.997)	94.5	95.8	550	0.971 (0.956–0.985)
Pancreas versus esophageal	0.999 (0.990–1.00)	97.1	98.9	550	0.960 (0.914–1.00)
Ovarian versus lung	0.994 (0.984–1.00)	97.6	96.6	600	1.00 (0.997–1.00)
Lung versus DLBCL	0.996 (0.992–0.999)	97.2	97.3	800	0.988 (0.983–0.993)
Sarcoma versus ovarian	0.995 (0.986–1.00)	99.2	95.7	300	1.00 (0.997–1.00)
Sarcoma versus GIST	1.00 (0.997–1.00)	99.3	100	300	1.00 (0.997–1.00)

AUC, area under the curve; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CLL, chronic lymphocytic leukemia; DLBCL, diffuse large B-cell lymphoma; GIST, gastrointestinal stromal tumor; LN, lymph node; MDS, myelodysplastic syndrome; MPN, myeloproliferative neoplasm.

Figure 1 Receiver operating characteristic curves for the prediction of diagnoses using targeted transcriptome according to artificial intelligence–based models. The area under the curve (AUC) and 95% CI are shown for various diagnostic classes. AML, acute myeloid leukemia; DLBCL, diffuse large B-cell lymphoma; FPF, false positive fraction (specificity); MDS, myelodysplastic syndrome; TPF, true positive fraction (sensitivity).

print & web 4C/FPO

FLA 5.6.0 DTD ■ AJPA3794_proof ■ 19 October 2022 ■ 10:22 am ■ EO: AJPA-D-22-

Table 3 Transcriptome and Differential Diagnosis Using Machine Learning Trained Using 47 Different Diagnostic Classes

Diagnosis	Cases, *N*	Accurate diagnosis as first choice (PPA), *n* (%)	PPV, %	Accurate diagnosis as second choice, *n* (%)	PPA by first and second 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

ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CLL, chronic lymphocytic leukemia; CML, chronic myeloid leukemia; CMML, chronic myelomonocytic leukemia; DLBCL, diffuse large B-cell lymphoma; MDS, myelodysplastic syndrome; MM, multiple myeloma; MPN, myeloproliferative neoplasm; PPA, positive percentage agreement; PPV, positive predictive value.

testing. Because some of the diagnostic classes had too few samples, the testing set included only 23 different diagnostic classes diagnosed against the 47 diagnostic classes (Table 3). To eliminate the underflow associated with the use of naïve bayesian classification, we applied the geometric mean to the likelihood produced by the naïve bayesian classifier. This allowed us to obtain a score for each diagnostic class. The score was used to rank the likelihood of each diagnosis. The purpose of this approach was to use additional information, particularly other clinical data, to select the best diagnosis.

The testing set included cases of hematologic and solid tumors. Evaluating each of the diagnostic classes individually showed variations between the diseases (Table 3). In acute lymphoblastic leukemia, 100% of the cases were correctly diagnosed, and the overall positive predictive value (PPV) was 84%. In contrast, none of the 17 chronic myeloid leukemia cases was correctly diagnosed, and most were classified as MPN or chronic myelomonocytic leukemia. As expected, all chronic myeloid leukemia cases were classified correctly when molecular abnormalities were assessed using the same RNA sequencing data. All chronic myeloid leukemia cases demonstrated the presence of breakpoint cluster region–abelson 1 (*BCR-ABL1*) fusion mRNA, and the PPV was 100%. There was a significant overlap in the diagnosis based on RNA expression alone between the normal BM, MDS, chronic myelomonocytic leukemia, MPN, and acute myeloid leukemia. Similarly, the BM was easily distinguished when the mutation profile was considered. The same was true when distinguishing between acute myeloid leukemia and MDS/chronic myelomonocytic leukemia. Mutation profiles were crucial for distinguishing between these myeloid entities.

The machine learning software correctly diagnosed 85% of diffuse large B-cell lymphoma cases, with a PPV of 73%. The diagnosis was 91% correct when the first and second choices were considered.

Among solid tumors, colorectal cancers were diagnosed correctly in 82% of the cases, with a PPV of 79%. Similarly, lung cancers were diagnosed correctly in 88% of the cases as the first choice, with PPVs of 73% and 91%, when both the first and second choices were considered (Table 3).

Sarcoma diagnosis was predicted in 65% of the cases, with a PPV of 45%. Most of the misdiagnosed cases of sarcoma were ovarian cancer and vice versa. This is most likely due to the presence of stromal elements in ovarian tumors, which was used for the training and diagnosis.

We also evaluated the diagnostic accuracy of this system by grouping the diagnostic classes into five groups: lymphoid, myeloid, carcinoma (including brain tumors), sarcoma, and normal (Table 4). As shown in Table 4, correct diagnosis was obtained in 84% of the cases as the first

UNCORRECTED PROOF

FLA 5.6.0 DTD ■ AJPA3794_proof ■ 19 October 2022 ■ 10:22 am ■ EO: AJPA-D-22-

ARTICLE IN PRESS

Table 4 Transcriptome and Differential Diagnosis between Five Major Diagnostic Classes Using Machine Learning and Targeted Transcriptome

Diagnosis	Cases, *N*	Cases correctly diagnosed as first choice (PPA), *n* (%)	Sensitivity (95% CI), %	Specificity (95% CI), %	Cases correctly diagnosed as second choice (PPA), *n* (%)	Cases correctly diagnosed as first and second choices (PPA), %
Lymphoid	427	389 (91)	77 (72–81)	88 (86–90)	20 (5)	96
Myeloid	295	258 (87)*	44 (38–49)	77 (75–80)	26 (9)	96
Carcinoma	452	427 (94)†	81 (77–84)	95 (92–96)	17 (4)	98
Normal	201	93 (46)‡	46 (39–53)	96 (95–97)	41 (20)	67
Sarcoma	40	26 (65)§	65 (48–79)	99 (98–99)	1 (3)	68
Total	1415	1189 (84)			109 (8)	92

*Of 37 patients, 36 were classified as normal.

†Of 25 patients, 14 were lymphoid because the tumor was metastatic to the lymph node or pleural fluid.

‡Of 108 cases, 107 were myeloid neoplasms (chronic myelomonocytic leukemia, myelodysplastic syndrome, chronic myeloid leukemia, or acute myeloid leukemia).

§Of 14 cases, 4 were classified as ovarian (Mullerian tumors) and 2 in lymph nodes.

PPA, positive percentage agreement.

ranked diagnostic choice, and an additional 8% as the second ranked diagnostic choice, with a final diagnostic accuracy of 92%. Most of the cases that were missed by the first choice and captured by the second choice had scores close to each other (<4%), indicating that the second option should be considered. Mostly, the missed cases were normal BM, most of which were misdiagnosed as MDS because of the significant similarity between the BM from MDS and normal BM. Particularly, all the normal BM samples were collected from patients having cytopenia and were considered negative for neoplasm because of lack of cytogenetic, morphologic, and molecular abnormalities. These cases were easily distinguished from MDS when mutation data were considered. Generally, misdiagnosed cases were commonly misdiagnosed within the same diagnostic category.

Discussion

Cancer is a genomic disease.[16] The DNA changes in cancer lead to a cascade of abnormalities in RNA expression, which, in turn, lead to abnormal phenotypes, including abnormal or lack of differentiation, uncontrolled growth and proliferation, and abnormal apoptosis.[17] Furthermore, these changes in cells trigger alterations in the host response that can be observed in the tumor microenvironment.[18] Analyzing the RNA of cancerous tissues can provide tremendous information on the biology of the tumor, its differentiation, and the surrounding microenvironment. This information provides insight into the clinical behavior and therapeutic efficacy of the tumors.[19] This information can be used to determine diagnosis, clinical course, potential therapeutic targets, and prognosis. In this study, we explored the potential of using RNA expression profiles to precisely diagnose cancer.

Determining the initial diagnosis is the first step in cancer management and therapy. Pathologists typically use the morphology and large panels of immunohistochemistry to confirm cancer diagnosis. Nevertheless, misdiagnosis may have a significant impact on clinical decisions and treatment, because morphologic evaluation is subjective and depends on the expertise of the pathologist. A more objective approach may help pathologists to diagnose precisely and reduce errors. This study investigated the potential of using RNA expression profiling in a machine learning approach to aid pathologists in making diagnoses and determining the cell of origin of the tumors. The approach described in this study was not intended to replace the clinical decision by the pathologist and clinician, but rather to aid the decision making and to add objectivity, efficiency, and reproducibility.

Although the focus of this article was to make a diagnosis and determine the cell of origin, the RNA molecular data generated in this process provided information on mutations, fusion genes, the microenvironment, and the immune response. As proof of principle, we focused on RNA expression profiling and did not incorporate mutation profiling in the diagnostic algorithms at this time. We used a targeted transcriptome to profile a wide range of hematologic neoplasms and solid tumors. We elected to use the targeted transcriptome rather than the whole transcriptome to exclude highly expressed housekeeping genes, and to improve the detection dynamic range of genes that may be expressed at low levels, which may have a significant impact on oncogenesis and cell differentiation. Furthermore, targeted transcriptome by hybrid capture is reliable when dealing with FFPE tissue and is more amenable to clinical testing and cost-effectiveness.

We first explored the potential of targeted RNA profiling combined with machine learning to distinguish between the two diagnostic classes. We used a unique approach in our

UNCORRECTED PROOF

FLA 5.6.0 DTD ■ AJPA3794_proof ■ 19 October 2022 ■ 10:22 am ■ EO: AJPA-D-22-

machine learning to select the proper genes for classification, as described in *Materials and Methods*. We elected to use machine learning over convolutional neural network or deep learning, which has been shown to be effective on image-related applications, because expression data are different. The design of convolutional neural networks, such as the multiple convolution layers with small windows, naturally fits the structure of image data. The grid structure of an image is effectively represented in convolutional neural networks. However, gene expression analysis is a different problem. Little is known on relations or structures among different genes.

Our approach is to combine the classifiers on individual genes without estimating their correlations. Estimating the mean and variance of one gene is certainly feasible, and our geometric mean naïve Bayes method provides a way to construct a stable and deterministic combined classifier. The risk of overfitting is low, because the parameter estimation is stable and there is no hyperparameter to fit. The power of our classifier comes from the contribution of a large number of single-gene classifiers. Each single-gene classifier is usually weak, but this weakness is overcome by combining many weak classifiers together, as we demonstrate in this article. We first ranked specific genes whose expressions could distinguish between the two diagnostic classes in question. We then used machine learning in combination with gene information to distinguish between the two classes. As shown in Table 2, Figure 1, and Supplemental Figure S1, we demonstrated that distinguishing between the two diagnostic classes was reliable. The accuracy of prediction can be calculated from the receiver operating characteristic curves and may vary dependent on which cutoff point is used and whether we want to emphasize sensitivity or specificity. For example, distinguishing between normal BM and BM in acute myeloid leukemia, and various types of leukemia, can be achieved with high sensitivity and specificity (>90%). As expected, distinguishing between normal BM and more chronic diseases was less conclusive (area under the curve of 78.1% for MDS and 90.9% for MPN) because of the overlap between these entities and normal or reactive BM. In particular, these BM samples were obtained because of some indication of abnormality but were determined to be negative for a neoplastic process by morphology, flow cytometry, and lack of mutations. Adding a mutation profile to the data and allowing the algorithm to consider mutations would significantly improve prediction. Another example is distinguishing between Hodgkin lymphoma and normal lymph nodes, which can be difficult based on morphology and immunoprobing by flow cytometry or immunohistochemistry. The machine learning algorithm was able to distinguish between these two diagnostic classes with high sensitivity and specificity (95.4% and 100%, respectively) using the expression profiles of 100 genes. Similarly, in solid tumors, distinguishing between two tumors based on the site of origin was achievable with high accuracy and area under the curve ranging between 1.00 and 0.959. As expected, distinguishing between endometrial cancer and ovarian cancer was relatively more challenging than distinguishing between other tumors (area under the curve = 0.959 using 600 genes). In solid tumors, the expression of many genes was required (between 300 and 900) to achieve high accuracy in predicting various diagnostic classes. In contrast, only 10 to 500 genes were needed to distinguish between various diagnostic classes of hematologic neoplasms.

When all the 47 diagnostic classes were considered and no prior knowledge of the tumor site and cell of origin was assumed, prediction of diagnosis solely based on RNA expression profiling was more challenging (Table 3). However, in this classification, the machine learning algorithm provided a ranking for potential diagnostic classes. This ranking system listed the biologically overlapping diagnostic classes; therefore, other information, including mutation profile, morphology, and clinical data, can be considered to reach the final diagnostic decision. The information obtained using this approach can be used in the two-class algorithm described above. As shown in Table 3, the positive percentage agreement was 100% for acute lymphoblastic leukemia and remained high for most of the major epithelial tumors (colorectal, brain, lung, and breast). The positive percentage agreement improved further when the second-ranking diagnosis was considered, particularly for lung cancer (88% to 91%). For hematologic neoplasms, high positive percentage agreement was obtained for lymphoid neoplasms, particularly diffuse large B-cell lymphoma, and improved further when a second-ranking diagnosis was considered (from 88% to 91% for diffuse large B-cell lymphoma, from 62% to 97% for Hodgkin lymphoma, and from 72% to 86% for follicular lymphoma). As expected, distinguishing between normal BM and chronic myeloid neoplasms was less reliable. Chronic myeloid leukemia was practically indistinguishable from other diseases without molecular data. However, the same targeted RNA sequencing provided the results of *BCR-ABL1* fusion mRNA, and the diagnosis could be confirmed. MDS, chronic myelomonocytic leukemia, MPN, and normal BM could be distinguished, if the mutation profile was considered.

By grouping samples into five classes (lymphoid, myeloid, carcinoma, normal, and sarcoma), we calculated the accuracy of diagnosis with sensitivity and specificity. As shown in Table 4, carcinomas and lymphomas were correctly classified with good sensitivity and specificity. Sarcoma and normal tissues were classified as having a high specificity.

This study demonstrated the potential of combining artificial intelligence with genomics in the routine practice of oncology, and in determining the diagnosis and cell origin of tumors. Therefore, clinical decisions can be based on solid objective data. However, some diagnostic classes

UNCORRECTED PROOF

ARTICLE IN PRESS

contained too few cases, and increasing the number of cases and further validation are needed. Some myeloid samples, particularly those of chronic leukemia, were in the early stage of disease. Samples with solid tumors were metastatic, involving lymph nodes. A metastatic epithelial tumor in a lymph node can show a lymphoid profile, in addition to carcinoma, particularly if the carcinoma fraction is not dominant. This could potentially confuse the diagnostic algorithm. Microdissection of the tumor was performed on all analyzed solid tumor samples, but the tumor fraction varied between 30% and 90%. Therefore, ranking the diagnoses was important, and it allowed us to consider other information to reach the final diagnosis. This approach was realized practically by developing a software that can be used to feed RNA data for automated diagnosis and classification of tumors. Furthermore, this software and algorithms can be continuously trained by adding more samples or new diagnostic classes.

Limitation of the study is not including the exact molecular mutations and chromosomal abnormalities in the algorithms. Integrating such abnormalities most likely will improve the prediction significantly.

Acknowledgment

We thank Editage for English-language editing.

Author Contributions

H.Z. and M.A. developed artificial intelligence algorithms and analyzed data; M.A.Q., M.W., and A.C. performed blind testing; A.E., A.I., J.M., M.D., D.S., M.G., A.P., A.G., and M.A. contributed samples, concept design, and data interpretation; and I.D.D., W.M., and I.S. performed RNA sequencing.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.ajpath.2022.09.006*.

References

1. Troyanskaya O, Trajanoski Z, Carpenter A, Thrun S, Razavian N, Oliver N: Artificial intelligence and cancer. Nat Cancer 2020, 1: 149−152
2. Adler-Milstein J, Chen JH, Dhaliwal G: Next-generation artificial intelligence for diagnosis: from predicting diagnostic labels to "wayfinding." JAMA 2021, 326:2467−2468
3. Elemento O, Leslie C, Lundin J, Tourassi G: Artificial intelligence in cancer research, diagnosis and therapy. Nat Rev Cancer 2021, 21: 747−752
4. Moon M, Nakai K: Stable feature selection based on the ensemble L 1-norm support vector machine for biomarker discovery. BMC Genomics 2016, 17:65−74
5. Hong M, Tao S, Zhang L, Diao L-T, Huang X, Huang S, Xie S-J, Xiao Z-D, Zhang H: RNA sequencing: new technologies and applications in cancer research. J Hematol Oncol 2020, 13:1−16
6. Govindarajan M, Wohlmuth C, Waas M, Bernardini MQ, Kislinger T: High-throughput approaches for precision medicine in high-grade serous ovarian cancer. J Hematol Oncol 2020, 13:1−20
7. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL: Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol 2012, 30:99−104
8. Reeser JW, Martin D, Miya J, Kautto EA, Lyon E, Zhu E, Wing MR, Smith A, Reeder R, Samorodnitsky E, Parks H, Naik KR, Gozgit J, Nowacki N, Davies KD, Varella-Garcia M, Yu L, Freud AG, Coleman J, Aisner DL, Roychowdhury S: Validation of a targeted RNA sequencing assay for kinase fusion detection in solid tumors. J Mol Diagn 2017, 19:682−696
9. Togni M, Masetti R, Pigazzi M, Astolfi A, Zama D, Indio V, Serravalle S, Manara E, Bisio V, Rizzari C, Basso G, Pession A, Locatelli F: Identification of the NUP98-PHF23 fusion gene in pediatric cytogenetically normal acute myeloid leukemia by whole-transcriptome sequencing. J Hematol Oncol 2015, 8:1−3
10. Veeraraghavan J, Ma J, Hu Y, Wang X-S: Recurrent and pathological gene fusions in breast cancer: current advances in genomic discovery and clinical implications. Breast Cancer Res Treat 2016, 158:219−232
11. Kloosterman WP, van den Braak RRC, Pieterse M, Van Roosmalen MJ, Sieuwerts AM, Stangl C, Brunekreef R, Lalmahomed ZS, Ooft S, Galen AV, Smid M, Lefebvre A, Zwartkruis F, Martens JWM, Foekens JA, Biermann K, Koudijs MJ, Ijzermans JNM, Voest EE: A systematic analysis of oncogenic gene fusions in primary colon cancer. Cancer Res 2017, 77:3814−3822
12. Zhou X, Zhan L, Huang K, Wang X: The functions and clinical significance of circRNAs in hematological malignancies. J Hematol Oncol 2020, 13:1−15
13. Liu Y, Cheng Z, Pang Y, Cui L, Qian T, Quan L, Zhao H, Shi J, Ke X, Fu L: Role of microRNAs, circRNAs and long noncoding RNAs in acute myeloid leukemia. J Hematol Oncol 2019, 12:1−20
14. Sun Y-M, Chen Y-Q: Principles and innovative technologies for decrypting noncoding RNAs: from discovery and functional prediction to clinical application. J Hematol Oncol 2020, 13:1−27
15. Albitar M, Zhang H, Goy A, Xu-Monette ZY, Bhagat G, Visco C, Tzankov A, Fang X, Zhu F, Dybkaer K, Chiu A, Tam W, Zu Y, Hsi ED, Hagemeister FB, Huh J, Ponzoni M, Ferreri AJM, Møller MB, Parsons BM, van Krieken JH, Piris MA, Winter JN, Li Y, Xu B, Young KH: Determining clinical course of diffuse large B-cell lymphoma using targeted transcriptome and machine learning algorithms. Blood Cancer J 2022, 12:25
16. Berger MF, Mardis ER: The emerging clinical relevance of genomics in cancer medicine. Nat Rev Clin Oncol 2018, 15:353−365
17. Curtius K, Wright NA, Graham TA: An evolutionary perspective on field cancerization. Nat Rev Cancer 2018, 18:19−32
18. Hanahan D: Hallmarks of cancer: new dimensions. Cancer Discov 2022, 12:31−46
19. Jarosz-Biej M, Smolarczyk R, Cichoń T, Kułach N: Tumor microenvironment as a "game changer" in cancer radiotherapy. Int J Mol Sci 2019, 20:3212

UNCORRECTED PROOF

FLA 5.6.0 DTD ■ AJPA3794_proof ■ 19 October 2022 ■ 10:22 am ■ EO: AJPA-D-22-