


ARTICLE



Multimetric feature selection for analyzing multicategory outcomes of colorectal cancer: random forest and multinomial logistic regression models

Catherine H. Feng¹, Mary L. Disis², Chao Cheng^{3,4,5} and Lanjing Zhang^{6,7,8,9} 

© The Author(s), under exclusive licence to United States and Canadian Academy of Pathology 2021

Colorectal cancer (CRC) is one of the most common cancers worldwide, and a leading cause of cancer deaths. Better classifying multicategory outcomes of CRC with clinical and omic data may help adjust treatment regimens based on individual's risk. Here, we selected the features that were useful for classifying four-category survival outcome of CRC using the clinical and transcriptomic data, or clinical, transcriptomic, microsatellite instability and selected oncogenic-driver data (all data) of TCGA. We also optimized multimetric feature selection to develop the best multinomial logistic regression (MLR) and random forest (RF) models that had the highest accuracy, precision, recall and *F1* score, respectively. We identified 2073 differentially expressed genes of the TCGA RNASeq dataset. MLR overall outperformed RF in the multimetric feature selection. In both RF and MLR models, precision, recall and *F1* score increased as the feature number increased and peaked at the feature number of 600–1000, while the models' accuracy remained stable. The best model was the MLR one with 825 features based on sum of squared coefficients using all data, and attained the best accuracy of 0.855, *F1* of 0.738 and precision of 0.832, which were higher than those using clinical and transcriptomic data. The top-ranked features in the MLR model of the best performance using clinical and transcriptomic data were different from those using all data. However, pathologic staging, *HBS1L*, *TSPYL4*, and *TP53TG3B* were the overlapping top-20 ranked features in the best models using clinical and transcriptomic, or all data. Thus, we developed a multimetric feature-selection based MLR model that outperformed RF models in classifying four-category outcome of CRC patients. Interestingly, adding microsatellite instability and oncogenic-driver data to clinical and transcriptomic data improved models' performances. Precision and recall of tuned algorithms may change significantly as the feature number changes, but accuracy appears not sensitive to these changes.

Laboratory Investigation (2022) 102:236–244; <https://doi.org/10.1038/s41374-021-00662-x>


INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer in both U.S. men and women when excluding melanomas¹. Genomic medicine and predictive markers have greatly helped treat CRC^{2,3}. We have shown that pathologic staging is critical for prognostication of CRC⁴. However, the 5-year overall survival of CRC was still only 64%⁵, and warrants identification of additional markers to better classify treatment responses and guide CRC treatments.

Transcriptomic and genomic data have been increasingly produced and used in recent decades⁶. However, there are several major challenges in integrating clinicopathologic and 'omics data for predicting clinical outcomes. Adding clinical data to transcriptomic data seems not to improve accuracy in classifying multicategory cancer outcomes⁷, but it is unknown whether adding genomic data of oncogenic drivers would help improve classification performance of statistical models. Moreover, limited by the outcome data in datasets, a vast majority of previous transcriptomic studies on CRC used binary or time-event

type outcomes^{8–16}, while clinicians need to predict clinical outcomes in much more details to better treat cancer patients and inform them of the disease prognosis. Furthermore, feature selection has greatly improved classification performances of various models on transcriptomic and microarray data^{17–20}. However, it is largely unclear whether addition of other 'omic data change the number of selected features. To explore these questions, we examined the performance metrics and optimal feature numbers of the best models in classifying four-category clinical outcomes of CRC before and after adding selected-genomic and microsatellite instability (MSI) data to clinical and transcriptomic data.

Machine learning (ML) algorithms produce a model that can perform classification, regression, and other similar tasks based on a given dataset, which can be used to predict the output of another system or dataset²¹. We have shown that transcriptomic data and clinical features can be used to accurately classify outcomes of lung, prostate, and breast cancers^{7,22–24}. However,

¹Montgomery High School, Skillman, NJ, USA. ²UW Medicine Cancer Vaccine Institute, University of Washington, Seattle, WA, USA. ³Department of Medicine, Section of Epidemiology and Population Sciences, Baylor College of Medicine, Houston, TX, USA. ⁴Department of Medicine, Baylor College of Medicine, Houston, TX, USA. ⁵Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA. ⁶Department of Biological Sciences, Rutgers University, Newark, NJ, USA. ⁷Department of Pathology, Princeton Medical Center, Plainsboro, NJ, USA. ⁸Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA. ⁹Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ, USA. email: lanjing.zhang@rutgers.edu

Received: 5 May 2021 Revised: 10 August 2021 Accepted: 12 August 2021

Published online: 18 September 2021

most of the model tuning was focused on accuracy, which is nonetheless very important. Additional metrics may be important to better understand the performance of an algorithm, particularly in imbalanced datasets^{25,26}. Hence, we aimed to understand the multimetric performance of random forests (RF) and multinomial logistic regression (MLR) in classifying multicategory outcomes of CRC patients, and how to select features for the best multimetric performances such as accuracy, precision, recall, and F1. We also compared the performance metrics of MLR and RF models and obtained the range of clinical and molecular features that are needed for best performance of each model.

METHODS

Data extraction

We obtained individual-level data of colorectal adenocarcinomas (pancancer atlas) from The Cancer Genome Atlas Program (TCGA) through cBioPortal repository (https://www.cbioportal.org/study/summary?id=coadread_tcga_pan_can_atlas_2018, Fig. 1)²⁷. The TCGA data were de-identified and publicly available. Therefore, this study was categorized as an exempt study (category 4) and did not require an Institutional Review Board (IRB) review. The transcriptomic data (RNAseq V2) were processed and normalized using RSEM (rsem.genes.normalized_results of the TCGA dataset), calculated into z-score and dichotomized using the median z-score of all samples²⁷.

We extracted the MSI statuses from a previous study on TCGA CRC that used a nonnegative matrix factorization-based consensus clustering model²⁸. We also extracted the genomic data of 16 clinically-significant oncogenic-driver or MSI-related genes, including *KRAS*, *NRAS*, *BRAF*, *NTRK1*, *NTRK2*, *NTRK3*, *ERBB2*, *POLE*, *MSH2*, *MSH6*, *PMS2*, *MLH1*, *APC*, *TP53*, *PIK3CA*, and *SMAD4*^{29–32}. The genomic data were standardized and annotated using Genome Nexus (which

utilizes VEP with the canonical UniProt transcript, https://github.com/mskcc/vcf2maf/blob/master/data/isoform_overrides_uniprot). Any driver alternations (amplifications, mutations, copy-number variation, and deletions) of these oncogenic driver genes were determined using GISTIC 2.0³³.

We performed the differentially expressed genes (DEG) analysis to remove the less relevant transcriptomes using Chi-square test⁷. The DEG and clinico-pathological features were subject to the modeling alone or with addition of the genomic and MSI data. The outcomes of the classification models were the patients' four-category survival, including alive with no progression, alive with progression, dead with no known progression, and dead with progression. All processes were conducted using Python 3.6.9.

RF modeling

Tuning: We used the RandomForestClassifier from the Python Scikit-learn package for RF modeling³⁴. We used the Gini index as the split criterion and had 20 iterations for each run.

During the tuning of the RF model, parameters including *n_estimator* (the number of trees in the forest), *min_sample_splits* (the minimum number of samples required to split an internal node), *min_samples_leaf* (the minimum number of samples required to be at a leaf node), and *n_jobs* (the number of jobs to run in parallel) were tuned.

To measure the accuracy of the model, we used *cross_val_score* imported from the Scikit-learn package, which evaluated the accuracy of a model using cross validation. We used a 5-fold cross validation, which produced five separate accuracy values for each iteration of our tuning, from which we took the average value.

We automated the tuning process. The pipeline first tuned *n_estimator* (range of 5–195, with increments of 5) and *min_sample_splits* (range of 2–14). This was repeated twice. We then tuned *min_sample_leaf* (range of 1–25) and *n_jobs* (range of 1–14) three times. The pipeline automatically selected the parameters which produced the highest average accuracy. To obtain the feature importance from our model, we used the *feature_importances_* property, which measured the Gini importance, the impurity-based feature importance. Feature importances was part of RandomForestClassifier from the Scikit-learn Package. We ran the tuned RF model 20 times to obtain the mean of the feature importance. We obtained performance metrics, including accuracy, recall, precision, and F1 score, using the *cross_validate* function imported from the *sklearn.model_selection* package from Scikit-learn³⁴.

Feature selection

Ranked approach, feature importance approach, and preset feature selection were used for feature selection, with each approach described in detail as following. The whole process of each approach was repeated 20 times. We then compared the performance metrics of the optimized models based on all data versus clinical and transcriptomic data that reached the best accuracy and precision, using Student t-test.

Ranked approach. We performed feature selection on the RF model using our pipeline based on ranking. After sorting the features based on their average feature importance, we created a number of reduced datasets. The number of features in these datasets increased by an increment of 185 features. Thus, we created 11 reduced datasets (Supplementary Table 1). During the process of feature selection, we tuned each reduced dataset individually using the aforementioned tuning process. After tuning, we ran the tuned model 20 times for each reduced dataset and collected the performance metrics of the model.

Feature importance approach. We performed feature selection using our pipeline based on the sum of the squared coefficients. Using the average feature importance which we previously obtained for each individual feature, we selected a number of sets of features using their feature importance values as a cut-off. We created 11 selected datasets based on the cut-off values (Supplementary Table 1). During the process of feature selection, we tuned each reduced dataset individually using the tuning process as described above. After tuning, we ran the tuned model 20 times for each reduced dataset and collected the performance metrics of the model.

Preset feature selection. We also performed feature selection using preset parameters for the RF model. Using the reduced datasets created for the feature importance approach, we set the values of *n_estimator* (range of 20–100, with increments of 20) and *min_sample_split* (range of 2–6, with

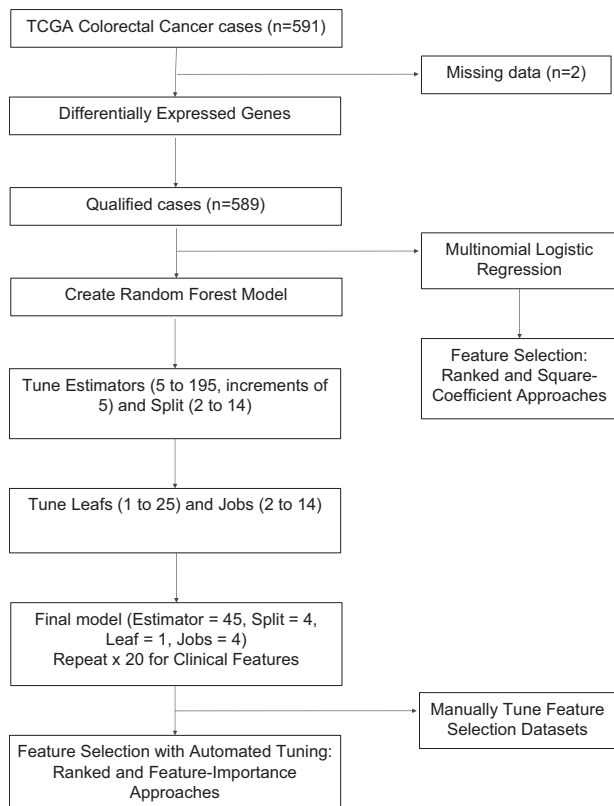


Fig. 1 Study flow. We extracted the colorectal adeno-carcinoma cases in the cancer genome atlas (TCGA), and classified the patient survival into four categories, including alive with no progression, alive with progression, dead with no known progression, and dead with progression. We used random forests (RF) and multinomial logistic regression (MLR) to classify the four-category outcomes. The 10-fold cross-validation approach was used during the modeling of the MLR model.

Table 1. Baseline characteristics of the included colorectal cancer cases by outcome.

Outcome	Alive no disease, N (%)	Alive with disease, N (%)	Dead no disease, N (%)	Dead with disease, N (%)
<i>Location</i>				
Colon	293 (72)	45 (70)	27 (77)	71 (85)
Rectal	113 (28)	19 (30)	8 (23)	13 (15)
<i>Age (years)</i>				
33–60	136 (33)	27 (42)	1 (3)	23 (27)
61–72	146 (36)	19 (30)	11 (31)	20 (24)
73+	124 (31)	18 (28)	23 (66)	41 (49)
<i>Sex</i>				
Female	196 (48)	28 (44)	22 (63)	33 (39)
Male	210 (52)	36 (56)	13 (37)	51 (61)
<i>Ethnicity</i>				
Unknown or not Hispanic	223 (55)	45 (70)	31 (89)	42 (50)
Hispanic	183 (45)	19 (30)	4 (11)	42 (50)
<i>Pathologic staging</i>				
1	90 (22)	6 (9)	5 (14)	2 (2)
2	162 (40)	24 (38)	12 (34)	21 (25)
3	112 (28)	20 (31)	14 (40)	24 (29)
4	42 (10)	14 (22)	4 (11)	37 (44)
<i>Pathologic T category</i>				
T1	18 (4)	1 (2)	1 (3)	1 (1)
T2	87 (21)	7 (11)	6 (17)	2 (2)
T3	269 (66)	46 (72)	24 (69)	60 (71)
T4	32 (8)	10 (16)	4 (11)	21 (25)
<i>Pathologic N category</i>				
N0	262 (65)	31 (48)	18 (51)	28 (33)
N1	97 (24)	14 (22)	7 (20)	23 (27)
N2	46 (11)	19 (30)	10 (29)	33 (39)
NX	1 (0)			
<i>Pathologic M category</i>				
M0	328 (82)	44 (69)	25 (76)	41 (49)
M1	33 (8)	13 (20)	1 (3)	35 (42)
MX	41 (10)	7 (11)	7 (21)	7 (8)
<i>Radiotherapy</i>				
NA	65 (16)	2 (3)	13 (37)	21 (25)
No	324 (80)	53 (83)	22 (63)	62 (74)
Yes	17 (4)	9 (14)		1 (1)
<i>Microsatellite instability</i>				
Stable	364 (90)	58 (91)	31 (89)	78 (93)
High	42 (10)	6 (9)	4 (11)	6 (7)

increments of 2) of the RF model uniformly for all datasets to investigate the nature and trends of dimension reduction in relation to tuning. The model was run with each combination of the *n_estimator* and *min_sample_split* values 20 times for all reduced datasets, and collected the performance metrics of the model.

MLR modeling

Model setup: We performed MLR to compare the results from our RF model with those from a more conventional model. We used LogisticRegression from the *sklearn.linear_model* package of *scikit-learn* to create the MLR model³⁴. MLR is a conventional statistical model, which doesn't need to be tuned. We used *train_test_split*, *cross_validate*, and *KFold* from the

sklearn.model_selection package to split the dataset into the test and train sets and perform a 10-fold cross validation. To extract the coefficient values for the features that built the MLR model, we used *coef_* from the *LogisticRegression* package.

Feature selection

Ranked approach. The pipeline using the ranked method was employed to perform the feature selection for the MLR model. Using the coefficients we obtained for each feature in a MLR, the pipeline sorted the features by their coefficients for each of the four classes of clinical outcomes in descending order. The pipeline then extracted the top and bottom features from every class to create reduced datasets and perform feature selection. We removed the duplicated features in each reduced dataset. To this end, we created 14 reduced datasets (Supplementary Table 1). During the process of performing Dimension Reduction, we ran the model 20 times for each reduced dataset and collected the performance metrics of the model.

Squared coefficient approach. Using the coefficient values, we obtained for each feature, the pipeline calculated the squared-coefficient of each feature by adding together the squared values of the feature's coefficient for each class. Then, these features were sorted in descending order based on their square-coefficient values. When creating the datasets, we selected square-coefficient values to create reduced datasets containing the features, whose square-coefficients were greater than the selected threshold. Using these thresholds, we created 10 reduced datasets (Supplementary Table 1). During the process of performing dimension reduction, we ran the model 20 times for each reduced dataset and collected the performance metrics of the model.

RESULTS

Baseline characteristics

There were 2034 DEGs among the 17,501 genes that were subject to RNAseq as shown by Chi-square test. The dataset had 589 cases, including 406 (68.7%) alive no disease, 65 (11.0%) alive with disease, 35 (5.9%) dead no disease, and 85 (14.4%) dead with disease (Table 1). There were 58 (9.9%) MSI-high CRC. The basic characteristics of these cases are summarized in Table 1.

Genomic alterations of CRC driver genes

Among the 16 most frequently mutated or clinically significant CRC driver genes, *APC*, *TP53* and *KRAS* were the most frequently seen, namely in 67% (393/589), 53% (312/589), and 38% (221/589) of the tumors, respectively (Supplementary Fig. 1 and Supplementary Table 2). The MSI related genes such as *MSH2*, *MSH6*, *PMS2*, *MLH1*, and *POLE* each had a mutation frequency of 3–6%, together accounting for about 8% of all tumors. The frequencies of driver-mutations in *KRAS G13D*, *KRAS Non-G13D*, *NRAS*, *BRAF V600E*, *PKI3A*, *APC*, and *TP53* were 9.3% (55/589), 28.2 (166/589), 5.1% (30/589), 8.2% (48/589), 22.9% (155/589), 64.9% (382/589), and 52.6% (310/589), respectively. Interestingly, *NTRK1*, *NTRK2*, and *NTRK3* driver mutations were seen in only 0.5% (3/589), 0% and 0% of the tumors, despite their 2–5% of overall mutation-frequencies. The most frequent genomic alteration of *ERBB2*, which is also clinically known as *HER2/Neu*, was amplification (3.4%, 20/589), including 2.9% (17/589) with amplification only and 0.5% (3/589) with amplification and mutation(s).

Tuning RF models

The RF model was tuned to find the best parameters to reach the highest possible accuracy. Through tuning, we found that the accuracy of the model can reach 69 or 70% when the parameters *n_estimator*, *min_sample_splits*, and *n_jobs* were not too few (i.e., *n_estimator* > 15, *min_sample_splits* > 2, *n_jobs* > 1), and the effect of *min_samples_leaf* to be less prominent when tuning (Fig. 2A, B).

The model was also tuned for its temporal efficiency. When the value of *n_estimator* was not too great (fewer than 130), the temporal efficiency of the model can be within 4 s (Fig. 3A). The

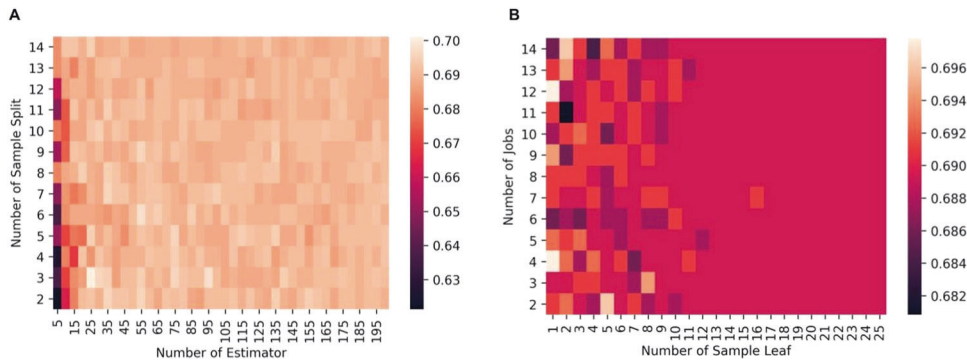


Fig. 2 Tuning the accuracy of the RF model. **A** Tuning the parameters `n_estimators` (from 5 to 195, in increments of 5) and `min_sample_splits` (from 2 to 14, in increments of 1). **B** Tuning the `min_samples_leaf` (from 1 to 25, in increments of 1) and `n_jobs` (from 2 to 14, in increments of 1).

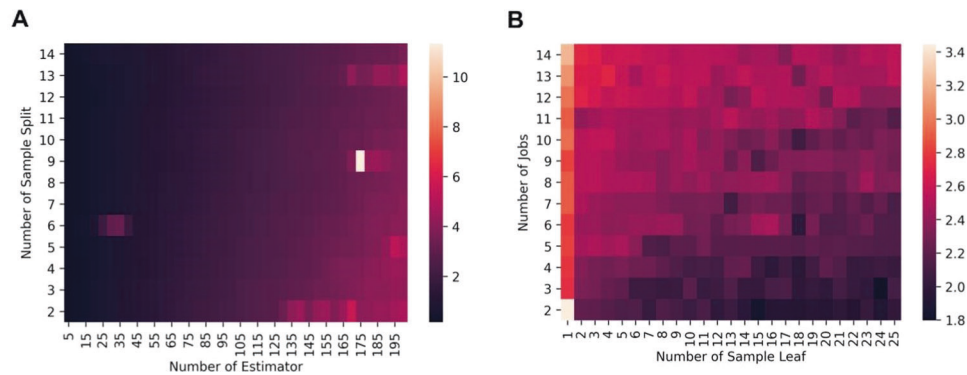


Fig. 3 Tuning the RF model's temporal efficiency. **A** Tuning the number of estimators and sample split. **B** Tuning the number of sample leaf and jobs without using graphic processing unit. The time of each individual run was measured in seconds.

effect of `min_sample_splits` was less prominent when tuning. Results also showed that when the value of `n_jobs` was sufficiently small (fewer than 10) and that of `n_samples_leaf` was sufficiently large (greater than 5), the efficiency of the model can range from 1.8 to 2.6 s (Fig. 3B).

Based on tuning results, the RF model was tuned with parameters `n_estimator` at 45, `min_sample_splits` at 4, `min_samples_leaf` at 1, and `n_jobs` at 4.

We performed 20 iterations of the tuned RF model, which appeared to have similar performances according to the confusion matrices and classification reports. Among these 20 iterations, the best classification accuracy of the model was 76% and it had a temporal efficiency of 2.57 s.

To evaluate the performance of a model, just looking at the accuracy and temporal efficiency is not enough. The values of precision, recall, and F1 all reflect the performance of the model in classifying each category. For our RF model, the precision, recall, and F1 in classifying alive with no progression were 76%, 99%, and 86%, respectively. However, the classification report also indicated that our model cannot effectively classify the other three categories (alive with progression, dead with no progression, and dead with progression).

The datasets used in this study had 2034 and 2051 features (also including MSI and 16 oncogenic-driver statuses), respectively. The individual importance values of these features were extracted and measured. The Gini importance of each individual feature was measured and averaged after 20 consecutive repeats using the tuned model.

Using the average feature importance values, we obtained the following important features from the clinical and transcriptomic data that had the top feature-importance values: `path_stage 4`, `path_n_stage 3`, `path_n_stage 1`, `COL11A2`, `SSX2IP`, `GSR`, `CYP46A1`,

`CCDC114`, `PSRC1`, `GABRD`, `GLT25D2`, `ASPDH`, `path_t_stage 4`, `ARHGAP4`, `CLK2`, `FLJ43663`, `TLE6`, `age_3grp3`, `HIST1H2BE`, `C20orf108`, `C15orf58`, `TP53TG3B`, `RBKS`, and `CD101`; and obtained the following from all data: `path_m_stage2`, `path_stage4`, `path_n_stage3`, `COL11A2`, `path_m_stage1`, `age_3grp3`, `C12orf69`, `path_n_stage1`, `CCDC114`, `CLK2P`, `FLJ12825`, `ASPDH`, `EFNA2`, `HIST1H2BE`, `path_t_stage4`, `ARHGAP4`, `CYP46A1`, `EVI5`, `CDHR2`, and `MMAA` (Supplementary Table 3).

During our manual tuning of the RF feature selection datasets based upon feature importance, we found that the number of features affected the parameters that led to the highest performance (Supplementary Fig. 2). For each of the manually tuned models, the datasets that underwent feature-selection performed better than the datasets with the highest number of features, which can be most clearly seen in the values of precision (Supplementary Fig. 2). These observations underscore the importance of feature selection before analysis to ensure its best performance.

The MLR using clinical and transcriptomic data (Supplementary Table 4) and all data (Supplementary Table 5) both showed that the features with higher association with class 1 (alive with no disease) usually had higher inversion association with classes 4 (dead with disease). Interestingly, some of the features were important in both datasets including `path_n_stage3`, `path_m_stage1`, `DCDC1`, `BRE`, and `TP53TG3B`.

Using the sum of squared coefficients, we obtained the following important features from the clinical and transcriptomic data: `path_stage 4`, `FLJ12825`, `KLF10`, `SLC5A5`, `path_n_stage1`, `GSR`, `SLC38A7`, `TUBA1C`, `DPCR1`, `TSPYL4`, `ERLIN2`, `HBS1L`, `path_stage2`, `LOC100130331`, `ZNF500`, `STOML1`, `TP53TG3B`, `TEX2`, `PPCDC`, `path_n_stage3`, `FAM124A`, `SLC20A2`, `EIF4E3`, and `SLC22A4`; and obtained the following from all data: `path_n_stage3`, `path_m_stage1`, `RRN3P2`,



Fig. 4 The accuracy remained relatively stable in some models using clinical and transcriptomic data, but showed peaks in others. Multi-metric feature selection for classifying four category outcomes of the colorectal cancers based on the random forest (RF) model (**A** based on ranking of feature importance; **B** based on feature importance) and multinomial logistic regression (MLR) model (**C** based on ranking of features' coefficients; **D** based on features' sum of squared coefficients) using clinical and transcriptomic data.

GPR88, *path_stage4*, *COMT*, *TP53TG3B*, *WBSCR17*, *BRE*, *GAS6*, *path_m_stage2*, *MSMP*, *FAM195B*, *GPRIN3*, *HBS1L*, *ZNF101*, *MOSC1*, *TSPYL4*, *TAS1R1*, and *KCNJ13* (Supplementary Table 6).

During our multimetric feature selection process on clinical and transcriptomic data (Fig. 4) and all data (Fig. 5), we found that accuracy remained relatively stable in RF models, while there were clear accuracy peaks in the MLR models. Interestingly, the number of features, that produced the best accuracy based on sum of squared coefficients using clinical and transcriptomic data, also produced the best precision and recall (Fig. 4D). Strikingly, the highest precision and recalls of MLR models using clinical and transcriptomic data after feature selection was 0.82 and 0.71, respectively, which were much higher than those in the RF model, while its highest accuracy was also higher than that of the RF model (0.85 versus 0.72) (Fig. 4). Similar findings were also noted in feature selection using all data (Fig. 5). These results indicate that the MLR model has better performance than the RF model in analyzing the datasets used in this study.

The best accuracy of all models and all datasets was achieved using MLR based on ranking of coefficients and 825 features of all data (0.855, Table 2), that also produced the highest precision

(0.832), F1 (0.738), and recall (0.698, tied with that of clinical and transcriptomic data). Adding selected genomic and MSI data slightly increased the best accuracy of each type of model and feature selection approach except MLR models based on sum of squared coefficients, as well as the best accuracy among all models (0.855 for all data versus 0.844 for clinical and transcriptomic data, $P < 0.001$, Table 2). Interestingly, the optimal feature numbers before and after adding selected genomic and MSI data seemed to be similar in all models except the RF model based on importance values.

DISCUSSION

In this study, we examined the clinical and molecular features that may be useful for classifying four-category survival outcome of CRC using the TCGA data. We created four pipelines which automated and optimized multimetric feature selection to identify the best model with the highest accuracy and precision. Our data show that the MLR models overall outperformed RF models in the multimetric feature selection. The best model was the MLR one with approximately 825 features based on sum of squared



Fig. 5 The accuracy remained relatively stable in some models using all data, but showed peaks in others. Multi-metric feature selection for classifying four category outcomes of the colorectal cancers based on the random forest (RF) model (A based on ranking of feature importance; B based on feature importance) and multinomial logistic regression (MLR) model (C based on ranking of features' coefficients; D based on features' sum of squared coefficients), using clinical, transcriptomic, microsatellite instability and selected genomic/oncogenic-driver data.

coefficients using all data, and attained the best accuracy of 0.855, F1 of 0.738 and precision of 0.832, which were better than those using clinical and transcriptomic data. A host of top features have been identified (Supplementary Tables 2 and 3). The top-ranked features in the MLR model of the best performance using clinical and transcriptomic data were different from those using all data. The pathologic stages, *HBS1L*, *TSPYL4*, and *TP53TG3B* were the overlapping top-20 ranked features in the best performing models using clinical and transcriptomic or all data.

We previously showed that, compared with transcriptomic data alone, combination of clinical and transcriptomic data did not significantly improve the accuracy of RF or MLR models for classifying multi-category lung adenocarcinomas⁷. However, adding MSI and selected genomic/oncogenic-driver data in this study significantly increased accuracy, precision, and F1 of the best feature-selection based MLR models, but not recall. The importance of alterations in MSI, *TP53*, *PIK3CA*, *BRAF*, and *SMAD4* has been shown before^{35–42}, and may explain the increase in model performance metrics. Indeed, the coefficients of *PMS2L4* (inverse

correlation), *TP53TG3B* (inverse correlation), *PIK3CA_DRIVER*, *BRAF_DRIVER*, and *SMAD4_DRIVER* were among the important features as assessed by the sum of squared coefficients. On the other hand, driver-alterations of *KRAS*, *NRAS*, *NTRK1*, *NTRK2*, *NTRK3*, *ERBB2*, and *APC* were not among the top- or bottom-500 features, despite their clinical values^{29–32}. One possible explanation is that no targeted therapies of these alterations were approved or widely used at the time when these TCGA CRC were treated. Such a time-difference was also noted for the utilization of radiotherapy. Interestingly, addition of MSI and selected oncogenic-driver data significantly increased accuracy, recall, and F1 of the best feature-selection based RF models, but not precision. This difference may be attributable to the different sensitivity of RF and MLR models to including more features in imbalanced datasets. However, future studies are warranted to examine the causes of these differences.

Despite the correlations in some top-ranked and bottom-ranked features, the top-ranked features in an optimized, tuned model may vary considerably by the databases and models'

Table 2. Performance metrics of the random forest and multinomial logistic regression models with the highest accuracy and precision that were identified using the multimetric feature selection approach.

Dataset	RF based on ranking of feature importance		P	RF based on feature importance value		P	MLR based on ranking of coefficients		P	MLR based on sum of squared coefficients		P	P for MLR based on ranking vs sum of squared coefficient)
	Clinical + RNAseq	All data		Clinical + RNAseq	All data		Clinical + RNAseq	All data		Clinical + RNAseq	All data		
Number of features	185	187		624	597		825	1032					
Accuracy (mean ± SD)	0.703 ± 0.004	0.711 ± 0.005	<0.001	0.708 ± 0.005	0.836 ± 0.007	0.001	0.855 ± 0.007	0.843 ± 0.005	<0.001	0.641	<0.001		
Precision (mean ± SD)	0.380 ± 0.028	0.411 ± 0.071	0.078	0.411 ± 0.051	0.804 ± 0.021	0.227	0.832 ± 0.020	0.822 ± 0.018	<0.001	0.619	0.106		
Recall (mean ± SD)	0.282 ± 0.005	0.297 ± 0.010	<0.001	0.306 ± 0.008	0.686 ± 0.022	<0.001	0.698 ± 0.017	0.678 ± 0.020	0.067	0.677	0.002		
F1 (mean ± SD)	0.263 ± 0.008	0.286 ± 0.016	<0.001	0.298 ± 0.011	0.717 ± 0.019	<0.001	0.738 ± 0.016	0.716 ± 0.019	0.001	0.552	<0.001		

All experiments were repeated 20 times ($n = 20$ for calculating mean and variances); RF, random forest. All data included the clinical, transcriptomic (RNAseq), genomic, and microsatellite instability data; P, calculated using two-sided Student's *t*-test. MLR multinomial logistic regression, SD standard deviation.

parameters^{7,22–24}. The same phenomenon was also observed in this study. Interestingly, pathologic staging, *HBS1L*, *TSPYL4*, and *TP53TG3B* have been ranked top 20 in the best models using clinical + transcriptomic and all datasets, respectively. They thus may be the truly important features for classifying clinical outcomes in CRC. This observation is partially supported by the known prognostic values of pathologic staging for CRC survivals^{4,43,44}. However, the literature review shows that *TSPYL4* is only reportedly important for prognosis of pancreatic and head and neck cancers^{45,46}, *HBS1L(-MYB)* is only found important for risks of hepatocellular carcinoma and development of myeloproliferative neoplasms^{47,48}, and *TP53TG3B* seems important only for the prognosis of uterine corpus endometrial carcinoma and hepatocellular carcinoma^{49,50}. Therefore, the prognostic values of *TP53TG3B*, *TSPYL4* and *HBS1L* for CRC appear novel. Strikingly, *TSPYL4* and *HBS1L* are both located on chromosome 6 (6q22.1 and 6q23.3, respectively, www.genecards.org) and may be closely related, while *TP53TG3B* is separately on 16p11.2. Future works may be focused on *TSPYL4* and *HBS1L* related drug development and prognostic markers.

Most of the published works on feature selection were based on LASSO^{51–53}, RF^{17,54–58}, support vector machine^{59–63}, and incremental recursive feature elimination^{64–67}. In this study, we for the first time compared the multimetric performances of RF and MLR models. Despite the possible overfitting, MLR clearly outperformed RF after proper feature selection in all four examined performance metrics. This will greatly enable us to rigorously classify patients with CRC or other cancer for proper treatments. It is particularly useful in the imbalanced datasets^{25,26}.

Another strength of this study is to analyze outcomes in four categories. Most, if not all, of the previous transcriptomic studies on CRC focused on binary outcomes or time-events^{8–16}. Thus, the multicategory survival outcomes are poorly understood in CRC patients, particularly regarding their associated clinical and transcriptomic features. Our findings will shed light on how to better manage and prognosticate CRC patients based on these features. Indeed, it is possible that CRC patients who died with disease may be under-treated or non-responsive to current treatment regimens, while the ones who died without disease may be over-treated or suffer other non-cancer causes. Therefore, our optimized models for four-category survival outcome will help stratify these patients and proper adjust their treatment regimens according to the predicted outcome. The specific applications of ML models, that are based on clinical and 'omic data, may include: 1. To choose the treatment option(s) so that the patient would stay alive without disease, or less preferably alive with disease; 2. To deescalate current treatment-regimen and increase dosages or treatment modalities for the patients who died without disease and those who died with disease, respectively, in future studies or practice; 3. To prepare the patients and their families with the predicted four-category outcomes, such as death with disease or alive with disease.

It should be mentioned that there are several limitations of this study. First, we do not have an external test set, although we rigorously examined the models using 5-fold cross-validation and repeated the run for 20 times. Such a limitation is inevitable due to the lack of high-quality large-scale 'omic studies with detailed patient outcomes. Second, the four-category outcome in this cohort is not very balanced since nearly two thirds of the patients were alive no disease at end of the follow-up period. Thus, the model may be biased to better predict these patients, and less powerful and less useful for the other categories of outcome (i.e., dead with or without disease, or alive with disease). Future large-scale studies with balanced-outcomes are warranted to examine our models and findings. Third, only one type of cancer is analyzed here. It will be interesting to test whether our findings hold true in other types of cancers. Fourth, adjuvant therapy data were not available in the TCGA, and thus not included in the

analyses despite their clinical values and our desire of including them. It is also noteworthy that radiotherapy in this study (TCGA cohort) was not used as often as nowadays. The main reason is that the included CRC were treated in the 2000s³¹, when radiotherapy was not used as widely as nowadays⁶⁸, possibly due to the lack of sufficient awareness, expertise, guidelines, and resources. Fifth, the outcomes of cancer recurrence, metastasis and specific causes of death will be more helpful than the four-category outcomes of this dataset, but not available in any large-scale 'omic datasets. Future works should focus on generating this kind of 'omic datasets. The TCGA dataset also has the limitation that scholars cannot assess its accuracy or explore further to fit special needs. Finally, sample size of the TCGA cohort may be too small despite its great values and wide use.

In summary, we here report a multimetric feature selection-based MLR model using all data that outperformed RF models in classifying four-category outcome of CRC patients. Adding MSI and selected genomic/oncogenic-driver data increased the performance metrics of RF and MLR models. We also showed the association of feature number with models' performance. Approximately 600–1000 clinical and features are needed to reach an optimal model with the highest accuracy, precision, and recall, while more or fewer features would lead to lower accuracy, precision, and recall. The model developed in this study has the potential to be used to analyze other types of cancers. The top important feature identified by the models we developed could help understand colorectal tumorigenesis and guide CRC treatments.

REFERENCES

- Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *Cancer J. Clin.* **71**, 7–33 (2021).
- Zhang, M. et al. Association of KRAS mutation with tumor deposit status and overall survival of colorectal cancer. *Cancer Causes Control* **31**, 683–689 (2020).
- Chavali, L. B. et al. Radiotherapy for patients with resected tumor deposit-positive colorectal cancer: a surveillance, epidemiology, and end results-based population study. *Arch. Pathol. Lab. Med.* **142**, 721–729 (2018).
- Mayo, E., Llanos, A. A., Yi, X., Duan, S. Z. & Zhang, L. Prognostic value of tumour deposit and perineural invasion status in colorectal cancer patients: a SEER-based population study. *Histopathology* **69**, 230–238 (2016).
- Siegel, R. L. et al. Colorectal cancer statistics, 2020. *Cancer J. Clin.* **70**, 145–164 (2020).
- Liu, D. D. & Zhang, L. Trends in the characteristics of human functional genomic data on the gene expression omnibus, 2001–2017. *Lab. Invest.* **99**, 118–127 (2019).
- Deng, F., Shen, L., Wang, H. & Zhang, L. Classify multicategory outcome in patients with lung adenocarcinoma using clinical, transcriptomic and clinico-transcriptomic data: machine learning versus multinomial models. *Am. J. Cancer Res.* **10**, 4624–4639 (2020).
- Sousa-Squiavinato, A. C. M. et al. Cofilin-1, LIMK1 and SSH1 are differentially expressed in locally advanced colorectal cancer and according to consensus molecular subtypes. *Cancer Cell Int.* **21**, 69 (2021).
- Zhang, Z. et al. Genomics and prognosis analysis of epithelial-mesenchymal transition in colorectal cancer patients. *BMC Cancer* **20**, 1135 (2020).
- Zhang, Z. et al. Comprehensive analysis of the transcriptome-wide m6A methylome in colorectal cancer by MeRIP sequencing. *Epigenetics* **16**, 1–11 (2020).
- Zhang, X. et al. Promoter hypermethylation of CHODL contributes to carcinogenesis and indicates poor survival in patients with early-stage colorectal cancer. *J. Cancer* **11**, 2874–2886 (2020).
- Tokunaga, R. et al. 12-Chemokine signature, a predictor of tumor recurrence in colorectal cancer. *Int. J. Cancer* **147**, 532–41 (2020).
- Saleh, R. et al. RNA-Seq analysis of colorectal tumor-infiltrating myeloid-derived suppressor cell subsets revealed gene signatures of poor prognosis. *Front. Oncol.* **10**, 604906 (2020).
- Ren, Y., Lv, Y., Li, T. & Jiang, Q. High expression of PLAC1 in colon cancer as a predictor of poor prognosis: a study based on TCGA data. *Gene* **763**, 145072 (2020).
- Poursheikhani, A., Abbaszadegan, M. R., Nokhandani, N. & Kerachian, M. A. Integration analysis of long non-coding RNA (lncRNA) role in tumorigenesis of colon adenocarcinoma. *BMC Med. Genomics* **13**, 108 (2020).
- Bala, P. et al. Exome sequencing identifies ARID2 as a novel tumor suppressor in early-onset sporadic rectal cancer. *Oncogene* **40**, 863–872 (2020).
- Moody, L., Chen, H. & Pan, Y. X. Considerations for feature selection using gene pairs and applications in large-scale dataset integration, novel oncogene discovery, and interpretable cancer screening. *BMC Med. Genomics* **13**, 148 (2020).
- Park, S. et al. Wx: a neural network-based feature selection algorithm for transcriptomic data. *Sci. Rep.* **9**, 10500 (2019).
- Momenzadeh, M., Sehhati, M. & Rabbani, H. A novel feature selection method for microarray data classification based on hidden Markov model. *J. Biomed. Inform.* **95**, 103213 (2019).
- Chiesa, M., Colombo, G. I. & Piacentini, L. DaMiRseq-an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification. *Bioinformatics* **34**, 1416–1418 (2018).
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015).
- Wang, J. et al. Predicting long-term multicategory cause of death in patients with prostate cancer: random forest versus multinomial model. *Am. J. Cancer Res.* **10**, 1344–1355 (2020).
- Deng, F. et al. Predict multicategory causes of death in lung cancer patients using clinicopathologic factors. *Comput. Biol. Med.* **129**, 104161 (2020).
- Deng, F., Huang, J., Yuan, X., Cheng, C. & Zhang, L. Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data. *Lab. Invest.* **101**, 430–441 (2021).
- Naseriparsa, M., Al-Shammari, A., Sheng, M., Zhang, Y. & Zhou, R. RSMOTE: improving classification performance over imbalanced medical datasets. *Health Inf. Sci. Syst.* **8**, 22 (2020).
- Jeni, L. A., Cohn, J. F. & De La Torre, F. Facing imbalanced data recommendations for the use of performance metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (Acii)*. 245–251 (IEEE Xplore, 2013).
- Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
- Hu, W. et al. Subtyping of microsatellite instability-high colorectal cancer. *Cell Commun. Signal.* **17**, 79 (2019).
- Benson, A. B. et al. Colon cancer, version 2.2021, NCCN clinical practice guidelines in oncology. *J. Natl Compr. Cancer Netw.* **19**, 329–359 (2021).
- Benson, A. B. et al. NCCN guidelines insights: rectal cancer, version 6.2020. *J. Natl Compr. Cancer Netw.* **18**, 806–815 (2020).
- Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
- Cocco, E., Scaltriti, M. & Drilon, A. NTRK fusion-positive cancers and TRK inhibitor therapy. *Nat. Rev. Clin. Oncol.* **15**, 731–747 (2018).
- Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
- Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Phipps, A. I. et al. Colon and rectal cancer survival by tumor location and microsatellite instability: the colon cancer family registry. *Dis. Colon Rectum* **56**, 937–944 (2013).
- Samowitz, W. S. et al. Microsatellite instability in sporadic colon cancer is associated with an improved prognosis at the population level. *Cancer Epidemiol. Biomark. Prev.* **10**, 917–923 (2001).
- Zhuang, Y. et al. Multi gene mutation signatures in colorectal cancer patients: predict for the diagnosis, pathological classification, staging and prognosis. *BMC Cancer* **21**, 380 (2021).
- Zhang, C. et al. microRNA-1827 represses MDM2 to positively regulate tumor suppressor p53 and suppress tumorigenesis. *Oncotarget* **7**, 8783–8796 (2016).
- Yan, P. et al. Reduced expression of SMAD4 is associated with poor survival in colon cancer. *Clin. Cancer Res.* **22**, 3037–3047 (2016).
- Voorneveld, P. W. et al. Loss of SMAD4 alters BMP signaling to promote colorectal cancer cell metastasis via activation of Rho and ROCK. *Gastroenterology* **147**, 196–208 e113 (2014).
- Ogino, S. et al. CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. *Gut* **58**, 90–96 (2009).
- Samowitz, W. S. et al. Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers. *Cancer Res.* **65**, 6063–6069 (2005).
- Washington, M. K. Colorectal carcinoma: selected issues in pathologic examination and staging and determination of prognostic factors. *Arch. Pathol. Lab. Med.* **132**, 1600–1607 (2008).
- Compton, C. C. & Greene, F. L. The staging of colorectal cancer: 2004 and beyond. *Cancer J. Clin.* **54**, 295–308 (2004).
- Xu, D. et al. Development and clinical validation of a novel 9-gene prognostic model based on multi-omics in pancreatic adenocarcinoma. *Pharmacol. Res.* **164**, 105370 (2021).

46. Pan, Y., Song, Y., Cheng, L., Xu, H. & Liu, J. Analysis of methylation-driven genes for predicting the prognosis of patients with head and neck squamous cell carcinoma. *J. Cell Biochem.* **120**, 19482–19495 (2019).
47. Kodama, T. et al. Two-step forward genetic screen in mice identifies Ral GTPase-activating proteins as suppressors of hepatocellular carcinoma. *Gastroenterology* **151**, 324–337 e312 (2016).
48. Tapper, W. et al. Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. *Nat. Commun.* **6**, 6691 (2015).
49. Liu, H., Li, H., Luo, K., Sharma, A. & Sun, X. Prognostic gene expression signature revealed the involvement of mutational pathways in cancer genome. *J. Cancer* **11**, 4510–4520 (2020).
50. Saelee, P. et al. Novel PNLIPRP3 and DOCK8 gene expression and prognostic implications of DNA loss on chromosome 10q25.3 in hepatocellular carcinoma. *Asian Pac. J. Cancer Prev.* **10**, 501–506 (2009).
51. Deshpande, S., Shuttleworth, J., Yang, J., Taramonli, S. & England, M. PLIT: An alignment-free computational tool for identification of long non-coding RNAs in plant transcriptomic datasets. *Comput. Biol. Med.* **105**, 169–181 (2019).
52. Jylhävä, J. et al. Identification of a prognostic signature for old-age mortality by integrating genome-wide transcriptomic data with the conventional predictors: the Vitality 90+ Study. *BMC Med. Genomics* **7**, 54 (2014).
53. Tolosi, L. & Lengauer, T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* **27**, 1986–1994 (2011).
54. Wang, J. & Wang, L. Prediction and prioritization of autism-associated long non-coding RNAs using gene expression and sequence features. *BMC Bioinform.* **21**, 505 (2020).
55. Ma, H., Tong, L., Zhang, Q., Chang, W. & Li, F. Identification of 5 gene signatures in survival prediction for patients with lung squamous cell carcinoma based on integrated multiomics data analysis. *Biomed. Res. Int.* **2020**, 6427483 (2020).
56. Lu, Z. et al. A 13-immune gene set signature for prediction of colon cancer prognosis. *Comb. Chem. High Throughput Screen.* <https://doi.org/10.2174/1386207323666200930104744> (2020)
57. Cheng, N., Schulte, A. J., Santosa, F. & Kim, J. H. Machine learning application identifies novel gene signatures from transcriptomic data of spontaneous canine hemangiosarcoma. *Brief Bioinform.* **22**, bbaa252 (2020).
58. Long, N. P. et al. High-throughput omics and statistical learning integration for the discovery and validation of novel diagnostic signatures in colorectal cancer. *Int. J. Mol. Sci.* **20**, 296 (2019).
59. Zhang, Z. Y. et al. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief. Bioinform.* **22**, 526–535 (2021).
60. Yuan, F., Lu, L. & Zou, Q. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochim. Biophys. Acta* **1866**, 165822 (2020).
61. Li, J. et al. Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine. *Cancer Gene Ther.* **27**, 56–69 (2020).
62. Fernández, E. A. et al. Unveiling the immune infiltrate modulation in cancer and response to immunotherapy by MIXTURE-an enhanced deconvolution method. *Brief. Bioinform.* **22**, bbaa317 (2020).
63. Chen, Z. et al. Feature selection may improve deep neural networks for the bioinformatics problems. *Bioinformatics* **36**, 1542–1552 (2020).
64. Mangiola, S. et al. Periprostatic fat tissue transcriptome reveals a signature diagnostic for high-risk prostate cancer. *Endocr. Relat. Cancer* **25**, 569–581 (2018).
65. Fatai, A. A. & Gamielien, J. A 35-gene signature discriminates between rapidly- and slowly-progressing glioblastoma multiforme and predicts survival in known subtypes of the cancer. *BMC Cancer* **18**, 377 (2018).
66. Hu, Y. et al. A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. *BMC Genomics* **17**, 1025 (2016).
67. Wei, X. et al. Identification of biomarkers that distinguish chemical contaminants based on gene expression profiles. *BMC Genomics* **15**, 248 (2014).
68. Murphy, C. C., Harlan, L. C., Lund, J. L., Lynch, C. F. & Geiger, A. M. Patterns of colorectal cancer care in the United States: 1990–2010. *J. Natl Cancer Inst.* **107**, djv198 (2015).

ACKNOWLEDGEMENTS

The work was in part supported by the Ramzi S. Cotran Young Investigator Award (to LZ) from the U.S. and Canadian Academy of Pathology. The funder plays no roles in the study design, data analysis or manuscript preparation.

AUTHOR CONTRIBUTIONS

C.H.F., C.C. and L.Z. designed the study, C.H.F. and L.Z. conducted the study and drafted the manuscript, all authors discussed, revised and edited the manuscript and L.Z. supervised the work.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41374-021-00662-x>.

Correspondence and requests for materials should be addressed to Lanjing Zhang.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.