Computational and Structural Biotechnology Journal 20 (2022) 2672–2679

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Can we infer tumor presence of single cell transcriptomes and their tumor of origin from bulk transcriptomes by machine learning?

Hua-Ping Liu [a,1], Dongwen Wang [a], Hung-Ming Lai [b,1,*]

[a] National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen 518116, China
[b] Aiphaqua Genomics Research Unit, Taipei 111, Taiwan

ARTICLE INFO

Article history:
Received 14 January 2022
Received in revised form 9 May 2022
Accepted 19 May 2022
Available online 23 May 2022

Keywords:
Single cell transcriptomes
Circulating tumor cells
RNA-seq
Translational bioinformatics

ABSTRACT

There is a growing need to build a model that uses single cell RNA-seq (scRNA-seq) to separate malignant cells from nonmalignant cells and to identify tumor of origin of single cells and/or circulating tumor cells (CTCs). Currently, it is infeasible to build a tumor of origin model learnt from scRNA-seq by machine learning (ML). We then wondered if an ML model learnt from bulk transcriptomes is applicable to scRNA-seq to infer single cells' tumor presence and further indicate their tumor of origin. We used k-nearest neighbors, one-versus-all support vector machine, one-versus-one support vector machine, random forest and introduced scTumorTrace to conduct a pioneering experiment containing leukocytes and seven major cancer types where bulk RNA-seq and scRNA-seq data were available. 13 ML models learnt from bulk RNA-seq were all reliable to use (F-score > 96%) shown by a validation set of bulk transcriptomes, but none of them was applicable to scRNA-seq except scTumorTrace. Making inferences from bulk RNA-seq to scRNA-seq was impaired by feature selection and improved by log2-transformed TPM units. scTumorTrace with transcriptome-wide 2-tuples showed F-score beyond 98.74 and 94.29% in inferring tumor presence and tumor of origin at single-cell resolution and correctly identified 45 single candidate prostate CTCs but lineage-confirmed non-CTCs as leukocytes. We concluded that modern ML techniques are quantitative and could hardly address the raised questions. scTumorTrace with transcriptome-wide 2-tuples is qualitative, standardization-free and not subject to log2-transformed quantities, enabling us to infer tumor presence of single cell transcriptomes and their tumor of origin from bulk transcriptomes.

1. Introduction

Single cell RNA-seq (scRNA-seq) is already a powerful biomedicine technique to study the tumor microenvironment (TME) [1], tumor diagnosis [2], and therapeutic resistance [3] and can be used to characterize circulating tumor cells (CTCs) [4]. For exploring the complexity of the TME, it is essential to effectively separate malignant cells from nonmalignant cells. The separation is equivalent to determining whether there is tumor present in single cells and a few tools have been developed for the separation by scRNA-seq [1,5,6].

Carcinoma of unknown primary origin (CUP) is a rare disease where malignant cells spread elsewhere in the body but routine testing cannot locate their origin [7]. CTCs have recently been attempted to find a starting point of occult primary cancer [8,9]. However, CTCs are cancer cells that shed from a primary or metastatic tumor lesion and circulate in the blood stream. The primary site the cancer began is usually known in current CTC clinical settings [10]; in fact, CTCs tumor of origin itself remains problematic if their tumor of origin is not given or not sure [11]. Using scRNA-seq to infer single cells' tumor of origin (scTOO) would be helpful in answering CUP and CTCs tumor of origin.

Unfortunately, scRNA-seq has its disadvantages of low capture efficiency and high dropouts [12,13]. The disadvantages can interfere with an accurate portrayal of single cell expression programs. scRNA-seq is currently costly and it is not economically feasible to capture single cell transcriptomes of thousands of patients over a wide range of cancer types for scTOO model training by machine learning (ML). Alternatively, bulk transcriptomes may be a practi-

* Corresponding author.
E-mail addresses: hung-ming.lai@outlook.com (H.-P. Liu), hmlai@aiphaqua.com, hung-ming.lai@outlook.com (H.-M. Lai).
[1] H.-M. Lai and H.-P. Liu contributed equally to this work.

https://doi.org/10.1016/j.csbj.2022.05.035

cal approach to learning a multiclass model for the scTOO inference.

Bulk RNA-seq detects an average of thousands of cells' gene expression and captures more transcripts with a lower level of technical noise than scRNA-seq [12,14]. It has been shown that 60% tumor purity is sufficient for a bulk tumor sample to represent a mass of malignant cells in the TME [15]. More importantly, public repositories have an abundant supply of bulk RNA-seq data ranging over plenty of tumor types with many hundreds of patients per type. Bulk RNA-seq and scRNA-seq are quite different on biological and technical levels. It is, therefore, worth knowing whether an ML multiclass model learnt from bulk transcriptomes can be applicable to scRNA-seq data to infer single cells' tumor presence and their tumor of origin.

2. Materials and methods

2.1. Experimental design

We carried out a pioneering experiment with seven major cancer types (ovary, lung, liver, colorectum, breast, prostate, and melanoma) and white blood cells (WBCs) sourced from two public repositories (TCGA and GEO). The tumors we used were all *in situ*, and the metastatic ones were discarded. WBCs were from health people and non-tumor patients who suffered from non-cancer diseases. We created a cohort of bulk RNA-seq samples and randomly split it into training and test (validation) sets in the ratio of 7:3 (Table S1). We used single cell RNA-seq data (ovary, lung, liver, colorectum, melanoma and WBCs), CTC RNA-seq data (breast and prostate) and lineage-confirmed non-positive CTC RNA-seq data (prostate) to form an application set (Table S2). scRNA-seq imputation was performed by the SCRABBLE algorithm [16]. We only considered single cell transcriptomic profiles whose genes were not preselected or not truncated. All of the single cells (CTCs included) were patient-derived, and single cells from cell lines were excluded.

We used four modern machine learning (ML) methods (kNN, one-versus-all SVM, one-versus-one SVM and random forest) and introduced a new learning technique (scTumorTrace) to build an ML multiclass model learnt from bulk training data. The built ML model was validated by bulk test data to show if it was reliable to use; if yes, we applied it to scRNA-seq data (Table S2). Both bulk RNA-seq and scRNA-seq data were normalized to the TPM (Transcripts Per Million) units as expression quantities. We examined standardization on two quantities: TPM and log2(TPM + 1). TPM (respectively, log2(TPM + 1)) values were mean-centered with a unit standard deviation and were termed z-score (respectively, standardized log2 transformation). Standardization was applied to each of a training, a test and an application set. We used the TPM units for scTumorTrace and the two sorts of standardized quantities for the four modern ML methods. We also examined whether feature selection impacts on the applicability to scRNA-seq. The examination was conducted in experiments of random forest and scTumorTrace.

2.2. Modern machine learning

We used k = 3 with Euclidean distance for k-nearest neighbors (kNN) and grew 100 trees to build random forest (RF). A linear kernel function with the penalty of C = 1 was utilized for support vector machine in one-versus-all (OvA SVM) and one-versus-one (OvO SVM) schemes.

2.3. A new learning technique: scTumorTrace

We suppose that there are K classes $\mathrm{C} = \{c_u\}_{u=1}^{K}$ to be studied in a training set $\mathrm{D} = \{d_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ where $n = \sum_u n_u$ and n_u is the number of samples d_i in c_u. $\forall d_i \in \mathrm{D}$, $y_i \in \mathrm{C}$ is a class label and $\boldsymbol{x}_i = [x_{i1} \cdots x_{ip}]$ is an expression vector profiled by a set of p genes $\mathscr{F}$. We define a transcriptomic 2-tuple $\omega = (h, k)$ between two classes $c_u \neq c_v \in \mathrm{C}$: given two genes $h \neq k \in \mathscr{F}$, $x_{ih} > x_{ik}$ (respectively, $x_{ih} < x_{ik}$) holds in at least 90 percent of samples d_i with $y_i = c_u$ (respectively, $y_i = c_v$). $\forall c_u, c_v \in \mathrm{C}$, we construct an entire set of m_{uv} 2-tuples between c_u and c_v and refer to it as $\Omega_{uv} = \{\omega_l\}_{l=1}^{m_{uv}}$. Upon the training set D, we connect all the Ω_{uv}s to build an all-in-one panel of transcriptomic 2-tuples for C, $\Omega = \{\Omega_{uv}^{r} | c_u, c_v \in \mathrm{C}, r = 1, \cdots, \binom{K}{2}\}$.

Given an unlabeled sample $\widehat{d} = (\boldsymbol{x}, \widehat{y})$ to be inferred, we define its discriminant score $f_v(u|\boldsymbol{x})$ to indicate that a class c_v predicts the likelihood of $\widehat{y}$ being c_u.

$$f_v(u|\boldsymbol{x}) = \frac{\sum_{\omega \in \Omega_{uv}} [x_h > x_k]}{m_{uv} - m_{uv}(\mathscr{F}_{\boldsymbol{x}}) - m_{uv}(\boldsymbol{x})} \tag{1}$$

In Eq. (1) $\mathscr{F}_{\boldsymbol{x}}$ is the gene set that profiles $\widehat{d}$, $m_{uv}(\mathscr{F}_{\boldsymbol{x}})$ is the number of invalid 2-tuples $\omega \in \Omega_{uv}$ whose gene (h or k) is undefined in $\mathscr{F}_{\boldsymbol{x}}$, and $m_{uv}(\boldsymbol{x})$ is the number of meaningless 2-tuples $\omega \in \Omega_{uv}$ having $x_h = x_k$. Following Eq. (1), we define $S(u|\boldsymbol{x})$ as an overall score of $\widehat{y}$ being c_u supported by the other classes.

$$S(u|\boldsymbol{x}) = (K-1)^{-1} \sqrt{\sum_{v \neq u} f_v(u|\boldsymbol{x}) \sum_{v \neq u} [f_v(u|\boldsymbol{x}) > 0.5]} \tag{2}$$

Consequently, $\widehat{y}$ is determined by $\widehat{y} = \underset{c_u \in \mathrm{C}}{\operatorname{argmax}} S(u|\boldsymbol{x})$. Note that $[Q]$ is an indicator function (alias the Iverson bracket) for a statement Q in Eq. (1) and Eq. (2).

2.4. Tumor presence inference

It is straightforward to infer single cells' tumor presence in terms of multiclass classification. A single cell is malignant, providing that a multiclass model identifies it as a cell from one of seven cancer types.

2.5. Feature selection

We applied random forest recursive feature elimination (RFRFE) [17] to a training set with an evaluation procedure of 10-fold cross-validation and an error rate as an evaluation measure to select a subset of 4608 genes for z-score (Fig. S1) and 1025 genes for standardized log2 transformation (Fig. S2). For scTumorTrace, we used a simple filter to reduce amounts of transcriptomic 2-tuples from a transcriptome-wide scale to a modest or a small scale (Table S3). Upon a training set, we set a threshold for any two classes $c_u, c_v \in \mathrm{C}$ and retained genes whose expression intensities above the threshold in>50% samples in both c_u and c_v. The retained genes between c_u and c_v were used to discover Ω_{uv}. All the thresholds and the amounts of used transcriptomic 2-tuples were in Table S4-S5.

2.6. Performance metrics

Let $n_{TP}, n_{TN}, n_{FP}, n_{FN}, n_P, n_N$ be the number of true positives, true negatives, false positives, false negatives, positive (malignancy), and negative cases, respectively. We used sensitivity (*TPR*), specificity (*TNR*), F-score ($F1$), accuracy (*ACC*) to evaluate the inference of tumor presence.

$$TPR = \frac{n_{TP}}{n_P},\ TNR = \frac{n_{TN}}{n_N}, F1 = \frac{2n_{TP}}{2n_{TP} + n_{FP} + n_{FN}}, ACC = \frac{n_{TP} + n_{TN}}{n_P + n_N}$$

For an evaluation of scTOO inferences, we let T_k, p_k, n_k be the number of correct predictions, predicted instances and actual instances for a class c_k, respectively and calculated its recall (R_k), precision (P_k) and F-score ($F1_k$).

$$R_k = \frac{T_k}{n_k}, P_k = \frac{T_k}{p_k}, F1_k = \frac{2R_kP_k}{R_k + P_k}$$

We then used macro recall (*maR*), macro precision (*maP*), macro F-score (*maF*1), micro accuracy (*miACC*) as summary statistics.

$$maR = \frac{\sum R_k}{K},\ maP = \frac{\sum P_k}{K}, maF1 = \frac{\sum F1_k}{K}, miACC = \frac{\sum T_k}{\sum n_k}$$

3. Results

The purpose of conducting the present study is to question whether a multiclass model learnt from bulk transcriptomes by machine learning can be applicable to single cell transcriptomes to infer single cells' tumor presence and their tumor of origin, graphically illustrated in Fig. 1A. Details of how we designed experiments to answer were given in section 2.1. We also developed a new learning technique (scTumorTrace) as a companion to our study. scTumorTrace was outlined in Fig. 1B. and was described in mathematical detail in section 2.3.

A validation set of bulk transcriptomic data showed that thirteen ML models learnt from bulk RNA-seq were all robust (F-score > 96%, Table 1). It indicated that the thirteen classifiers were all eligible to be examined on scRNA-seq data for their applicability to single cells' inferences. All the modern ML techniques were unable to discriminate leukocytes from neoplastic cells (single tumor cells and circulating tumor cells) while three scTumorTrace classifiers were able to (Table 2). Their discriminating power increased with a growing number of employed transcriptomic 2-tuples (sensitivity, 97.24, 98.35 and 97.77%, and specificity, 65.10, 84.82 and 99.79%; Table 2 and Table S3-S5).

OvA SVM built on standardized log2-trandformed quantities demonstrated limited effectiveness (F-score = 57.99%) in inferring single cells' tumor of origin (scTOO) and the other nine modern ML models were not applicable to the inference (Table 3). Compared to z-score, standardized log2 transformation helped a quantitative ML classifier learnt from bulk transcriptomes address

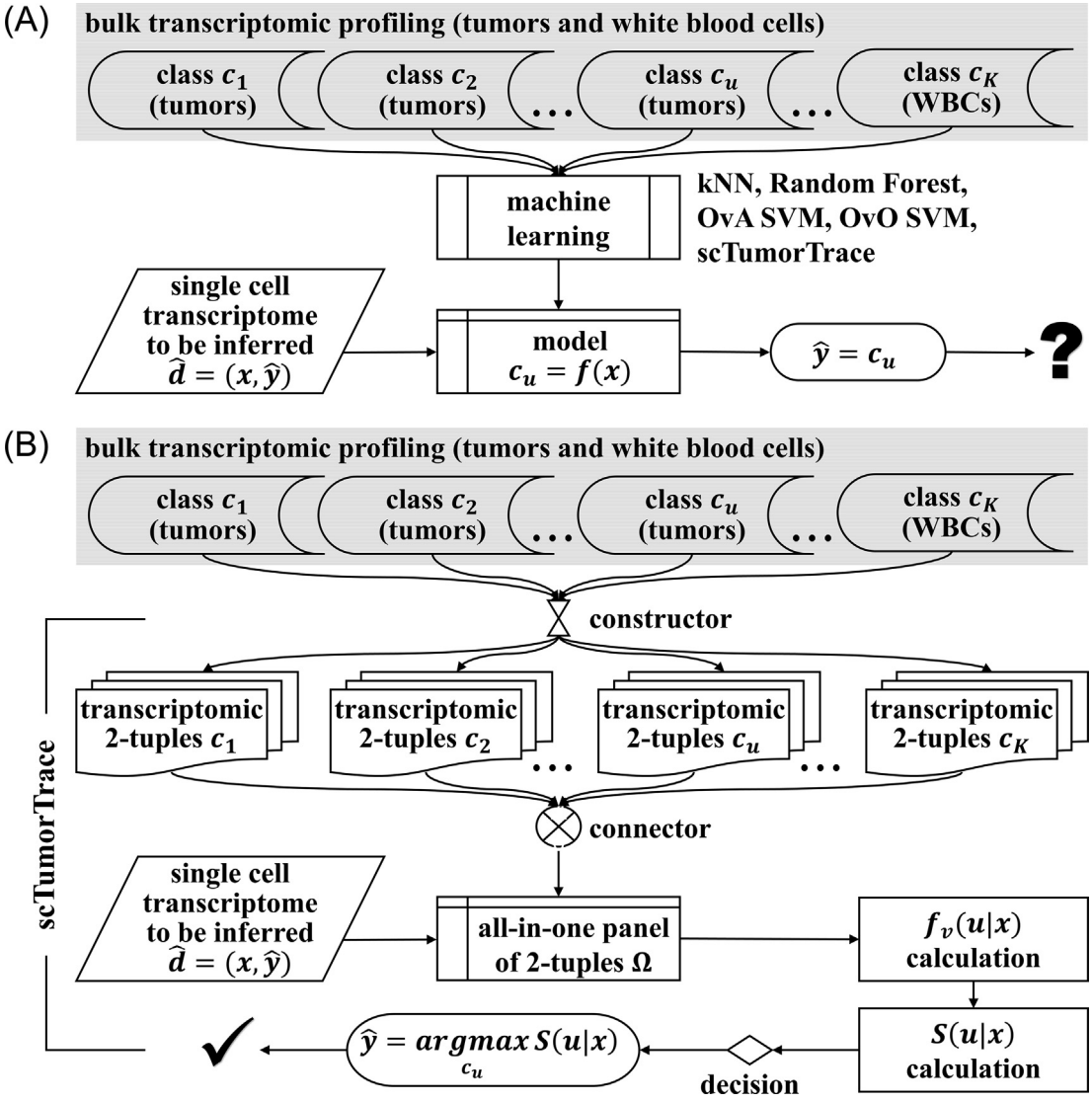

Fig. 1. Graphic outlines. (A) The purpose of the present study. (B) A new learning technique of scTumorTrace.

Table 1
Summary statistics of predictive performance: bulk tissue of origin (validation).

ML Methods	Feature Type	FS	Z	Log2	#Feat	R	P	F1	ACC	Effective
kNN	Quantitative	N	Y	N	13,126	96.02	96.98	96.39	96.94	✔
				Y		99.39	99.23	99.30	99.47	✔
OvA SVM	Quantitative	N	Y	N	13,126	99.47	99.31	99.38	99.58	✔
				Y		99.76	99.81	99.78	99.79	✔
OvO SVM	Quantitative	N	Y	N	13,126	99.12	99.54	99.32	99.37	✔
				Y		99.67	99.77	99.72	99.68	✔
RF	Quantitative	N	Y	N	13,126	99.37	99.15	99.25	99.26	✔
				Y		99.68	99.55	99.62	99.68	✔
RF	Quantitative	Y	Y	N	4608	99.52	99.40	99.46	99.47	✔
				Y	1025	99.61	99.31	99.45	99.58	✔
scTumorTrace	Qualitative	Y	N	N/Y	500+	97.51	97.65	97.55	97.36	✔
scTumorTrace	Qualitative	Y	N	N/Y	7 K+	98.48	98.33	98.39	98.31	✔
scTumorTrace	Qualitative	N	N	N/Y	200 K+	99.04	98.62	98.80	99.05	✔

kNN = k-nearest neighbors, OvA SVM = one-versus-all support vector machine, OvO SVM = one-versus-one support vector machine, RF = random forest, FS = feature selection is used (Y) or not (N), Z = standardization (z-score) is used (Y) or not (N), Log2 = a log2 scale is used (Y) or not (N), #Feat = amount of features, 500+=500 features (2-tuples) on average, 7 K+=7000 features (2-tuples) on average, 200 K+=0.2 million features (2-tuples) on average, R = recall%, P = precision%, F1 = f-score%, ACC = accuracy%, ✔= highly effective.

Table 2
Summary statistics of predictive performance: tumor presence of single cell transcriptomes.

ML Methods	Feature Type	FS	Z	Log2	#Feat	TPR	TNR	F1	ACC	Effective
kNN	Quantitative	N	Y	N	13,126	99.95	0	61.20	44.09	×
				Y		100	0	61.22	44.12	×
OvA SVM	Quantitative	N	Y	N	13,126	100	0	61.22	44.12	×
				Y		99.52	0.13	61.05	43.98	×
OvO SVM	Quantitative	N	Y	N	13,126	100	0	61.22	44.12	×
				Y		100	0	61.22	44.12	×
RF	Quantitative	N	Y	N	13,126	100	0	61.22	44.12	×
				Y		100	0	61.22	44.12	×
RF	Quantitative	Y	Y	N	4608	100	0	61.22	44.12	×
				Y	1025	100	0	61.22	44.12	×
scTumorTrace	Qualitative	Y	N	N/Y	500+	97.24	65.10	80.55	79.28	Δ
scTumorTrace	Qualitative	Y	N	N/Y	7 K+	98.35	84.82	90.40	90.79	▲
scTumorTrace	Qualitative	N	N	N/Y	200 K+	97.77	99.79	98.74	98.90	✔

kNN = k-nearest neighbors, OvA SVM = one-versus-all support vector machine, OvO SVM = one-versus-one support vector machine, RF = random forest, FS = feature selection is used (Y) or not (N), Z = standardization (z-score) is used (Y) or not (N), Log2 = a log2 scale is used (Y) or not (N), #Feat = amount of features, 500+=500 features (2-tuples) on average, 7 K+=7000 features (2-tuples) on average, 200 K+=0.2 million features (2-tuples) on average, TPR = sensitivity%, TNR = specificity%, F1 = f-score%, ACC = accuracy%, ×= not effective, Δ = less effective, ▲= fairly effective, ✔= highly effective.

Table 3
Summary statistics of predictive performance: tumor of origin of single and/or circulating tumor cells.

ML Methods	Feature Type	FS	Z	Log2	#Feat	R	P	F1	ACC	Effective
kNN	Quantitative	N	Y	N	13,126	38.31	NaN	NaN	9.21	×
				Y		61.60	NaN	NaN	29.58	×
OvA SVM	Quantitative	N	Y	N	13,126	63.77	NaN	NaN	14.63	×
				Y		80.96	57.50	57.99	42.80	Δ
OvO SVM	Quantitative	N	Y	N	13,126	51.96	NaN	NaN	12.52	×
				Y		71.83	NaN	NaN	39.55	×
RF	Quantitative	N	Y	N	13,126	47.96	NaN	NaN	12.40	×
				Y		67.20	NaN	NaN	41.21	×
RF	Quantitative	Y	Y	N	4608	38.39	NaN	NaN	8.98	×
				Y	1025	49.14	NaN	NaN	16.67	×
scTumorTrace	Qualitative	Y	N	N/Y	500+	63.60	54.46	55.57	74.89	Δ
scTumorTrace	Qualitative	Y	N	N/Y	7 K+	75.60	NaN	NaN	87.79	Δ
scTumorTrace	Qualitative	N	N	N/Y	200 K+	91.96	97.38	94.29	98.57	✔

kNN = k-nearest neighbors, OvA SVM = one-versus-all support vector machine, OvO SVM = one-versus-one support vector machine, RF = random forest, FS = feature selection is used (Y) or not (N), Z = standardization (z-score) is used (Y) or not (N), Log2 = a log2 scale is used (Y) or not (N), #Feat = amount of features, 500+=500 features (2-tuples) on average, 7 K+=7000 features (2-tuples) on average, 200 K+=0.2 million features (2-tuples) on average, R = recall%, P = precision%, F1 = f-score%, ACC = accuracy%, ×= not effective, Δ = less effective, ✔= highly effective.

scTOO (see accuracy, Table 3). The five experiments of random forest and scTumorTrqace showed that feature selection caused severe damage to the scTOO inference from bulk transcriptomes no matter what expression quantities were used (Table 3). scTumorTrace with transcriptome-wide 2-tuples (i.e. without performing feature selection) was the only classifier that was well able to infer single cells tumor presence (F-score = 98.74%, Table 2) and their tumor of origin (F-score = 94.29%, Table 3 and Fig. 2C).

Fig. 2D showed that 45 single candidate prostate CTCs but lineage-confirmed non-CTCs (false CTCs) were all accurately identified as leukocytes by scTumorTrace. Mirrored histograms further showed that almost all the false CTCs were identified unequivo-

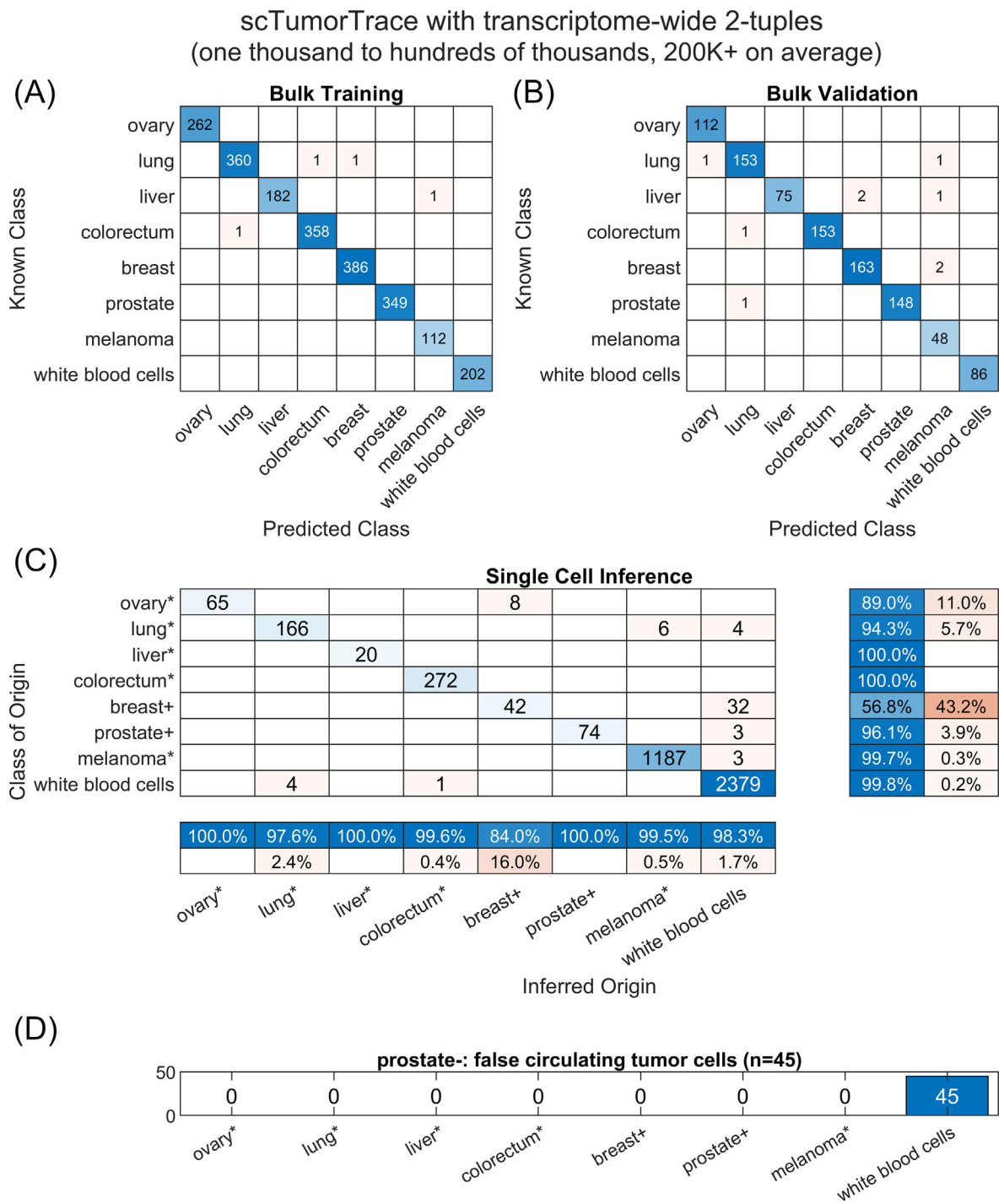

Fig. 2. Confusion matrix of scTumorTrace with transcriptome-wide 2-tuples. (A) Bulk tissue of origin: training set. (B) Bulk tissue of origin: validation set. (C) Tumor of origin of single cells and circulating tumor cells. (D) Inference of non-positive CTCs derived from prostate cancer patients.

cally (Fig. 3). Although false-CTC 45 was not wrongly identified, its inference was not very strongly supported by the other six cancer types, i.e. no pale-red bars against blue bars among them (Fig. 3). Its scRNA-seq profiling might be badly distorted and would not be very dissimilar to that of prostate cancer; however, the slight difference could still be detected by scTumorTrace. 3 out of 77 lineage-confirmed single prostate CTCs were also identified as leukocytes. The three circulating cells (cell 3, 21 and 24) had similar mirrored histograms to those of false-CTC 19, 20 and 37 (Fig. 3) so their tumor presence might be in doubt. Overall, scTumorTrace had both AUROC and AUPRC far beyond 90% in scTOO inferences except the breast cancer experiment (Fig. S3). 74 single breast CTCs were identified as either breast-derived malignant cells (n = 42) or leukocytes (n = 32) that resulted in a 98.04% AUROC and a 69.65% AUPRC (Fig. 2C and Fig. S3).

4. Discussion

CTC detection platforms normally require enrichment strategies and might not avoid false-positive or false-negative events [18,19].

Fig. 3. Mirrored histograms of discriminant scores for non-positive CTCs captured from prostate cancer patients. A mirrored histogram showed the likelihood of a single cell being prostate cancer-derived (the top histogram in red) or being leukocytes-like (the bottom histogram in blue) supported by a third-party cancer type. Pale-red indicated a cell to be inferred was against a single prostate CTC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

74 out of 77 lineage-confirmed single prostate CTCs were correctly inferred (Fig. 2C), and their discriminant scores showed overwhelming odds in favor of malignant cells derived from prostate cancer (red bars vs. blue & pale-blue bars; Fig. S4). Although the other three single prostate CTCs (cell 3, 21 and 24) seemed misclassified (as leukocytes), their discriminant scores were more similar to the 45 single candidate prostate CTCs but lineage-confirmed non-CTCs (blue bars vs. red & pale-red bars; Fig. 3). We, thus, doubted whether the three were false-positive events (i.e. tumor absence). In 74 single breast CTCs inferences, Fig. S5 (inferred leukocytes) and Fig. S6 (inferred breast-derived CTCs) showed that there were no overwhelming discriminant scores for breast CTCs (red bars) against leukocytes (blue bars). Here, red or pale-red bars indicated whether scTumorTrace tended towards or tended against breast cancer cells shaped by the other cancer types. Similarly, blue or pale-blue bars indicated white blood cells delineation. 32 single

breast CTCs were falsely identified as leukocytes, mainly because their discriminant scores between breast cancer and WBCs tended towards WBCs (pale-red against blue bars, Fig. S5). 40 out of 42 correctly inferred breast-derived CTCs also showed pale-red vs. blue bars in breast cancer cells vs. white blood cells (Fig. S6). Fig. S5 and Fig. S6 imply that discovered 2-tuples between breast cancer and WBCs would not be applicable to single breast cancer cell transcriptomes. Even so, Eq. (2) show that scTumorTrace can still draw their inferences to a certain extent by transcriptomic 2-tuples between breast cancer and the other six cancer types. Cell 44 had overwhelming odds against single breast CTC with the majority of pale-red bars (including overall score) so it would be a false positive rather than a misclassified instance (Fig. S5). Although accuracy did not reach a high level, scTumorTrace identified 74 single breast CTCs as either breast-derived malignant cells (n = 42) or leukocytes (n = 32) - nothing else but the two classes (Fig. 2C). Provided that tumor presence can be (or has been) experimentally confirmed, scTumorTrace can simply be used to answer CTCs tumor of origin with scRNA-seq profiling. Otherwise scTumorTrace can be an alternative remedy for the true identity of single CTCs synchronized with their tumor of origin inference.

Our empirical study showed that scTumorTrace was capable of inferring single cells' tumor presence (sensitivity, 97.77%, positive predictive value, 99.73%, and F-score, 98.74%). scTumorTrace can, therefore, help separate neoplastic cancer cells from the TME where tumor-infiltrating lymphocytes are present and interact closely with surrounding tumor cells. A recent method, CopyKAT, uses scRNA-seq data to infer aneuploid copy number events for the separation. However, it is not suitable for those cancers with few copy number alternations (CNA) and has biased detection of CNA events provided that the data to be inferred has a complete absence of tumor cells [6]. scTumorTrace employs transcriptome-wide 2-tuples that can be discovered among all sorts of cancers so it is also applicable to pediatric cancers and hematopoietic cancers for which CopyKAT is not suitable. More importantly, CopyKAT needs a bunch of scRNA-seq data for CNA inference while scTumorTrace is a classifier of single-instance inference and infers cells one by one no matter whether a tumor cell is absent or present in a scRNA-seq dataset. Briefly, scTumorTrace can be applicable to even one single cell but CopyKAT cannot.

Bulk RNA-seq estimates global expression of thousands of cells and can capture more transcripts while scRNA-seq detects an individual cell's expression with low capture efficiency and exhibits technical & biological cell-to-cell variation [20]. The present study observed varying levels of expression quantification between bulk RNA-seq and scRNA-seq (IQR: bulk training, 10.3495, bulk test, 10.3943, and single cells, 32.5449). This might explain why log2-transformed quantities (IQR: bulk training, 2.8230, bulk test, 2.8157, and single cells, 5.0493) were of help for quantitative ML techniques to infer scRNA-seq from bulk RNA-seq (Table 3). Since gene programs learnt from bulk transcriptomic identities might be distorted in single cell transcriptomes profiled by current scRNA-seq technologies, inferring single cells' tumor presence and their tumor of origin from bulk RNA-seq is a big challenge. Performing feature selection would make the challenge more challenging no matter the quantitative or the qualitative approaches. The fewer the features were selected; the more chance the tumor identities were damaged (Table 2-3). scTumorTrace can automatically adjust transcriptomic 2-tuples for each individual cell in accordance with its completeness of scRNA-seq profiling (see Eq. (1)). Therefore, scTumorTrace maximizes its effectiveness only when transcriptome-wide gene programs are available (Table 2-3 and Fig. 2C). scTumorTrace is a qualitative and a standardization-free learning technique and is not subject to log2-transformed quantities such that it can address the raised questions by "qualitative identities" inherent in both bulk transcriptomes and single cell transcriptomes; conversely, modern quantitative ML techniques cannot.

scTumorTrace has a fundamental weakness in computation time. This is due to a large number of transcriptome-wide 2-tuples ranging from one thousand to hundreds of thousands. When we extend the present study to more cancer types, an all-in-one panel of transcriptomic 2-tuples Ω can grow rapidly. A faster version should be developed, especially, for a high-throughput single cell platform like 10X Genomics. Meanwhile, we will need to improve the distinction between single breast cancer cells and white blood cells. We will also have to apply scTumorTrace to a broad range of cancer types such that we may suggest the primary site of the CUP disease by either bulk tissue or single cell transcriptomes.

5. Conclusions

Bulk RNA-seq and scRNA-seq are quite different on biological and technical levels. We questioned whether a multiclass model learnt from bulk RNA-seq is applicable to addressing single cells' tumor presence and their tumor of origin (scTOO). Our pioneering experiment produced three pieces of empirical evidence. Firstly, standardized log2 transformation is helpful to a quantitative ML method in improving its applicability. Secondly, performing feature selection causes damage to the applicability no matter the quantitative or the qualitative ML approaches. Thirdly, it is unlikely that we infer tumor presence of single cell transcriptomes and scTOO from bulk transcriptomes by modern quantitative machine learning. We might need to seek a qualitative learning technique with transcriptome-wide gene programs for such inferences. scTumorTrace could then be tailored to the particular needs.

Conflict of interest statement

Hung-Ming Lai has ownership interests (including stock, patents, etc) as an inventor of pending unpublished provisional patent application(s) for scTumorTrace and its clinical applications. No potential conflicts of interest were disclosed by the other authors.

Acknowledgments

In memory of the late Professor Zheng Guo, who made a positive contribution to the initial stages of conceptualization for this work.

Funding information

Dongwen Wang is supported by Shenzhen High-level Hospital Construction Fund and Sanming Project of Medicine in Shenzhen (No. SZSM202111003).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.05.035.

References

[1] Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014;344:1396–401.
[2] Gawel DR, Serra-Musach J, Lilja S, Aagesen J, Arenas A, Asking B, et al. A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases. Genome Med 2019;11:47.
[3] Sharma A, Cao EY, Kumar V, Zhang X, Leong HS, Wong AML, et al. Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy. Nat Commun 2018;9:4931.

[4] Miyamoto DT, Zheng Y, Wittner BS, Lee RJ, Zhu H, Broderick KT, et al. RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science 2015;349:1351–6.
[5] Fan J, Lee HO, Lee S, Ryu DE, Lee S, Xue C, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res 2018;28:1217–27.
[6] Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol 2021;39:599–608.
[7] Varadhachary GR. Carcinoma of unknown primary origin. Gastrointest Cancer Res 2007;1:229–35.
[8] Lu SH, Tsai WS, Chang YH, Chou TY, Pang ST, Lin PH, et al. Identifying cancer origin using circulating tumor cells. Cancer Biol Ther 2016;17:430–8.
[9] Matthew EM, Zhou L, Yang Z, Dicker DT, Holder SL, Lim B, et al. A multiplexed marker-based algorithm for diagnosis of carcinoma of unknown primary using circulating tumor cells. Oncotarget 2016;7:3662–76.
[10] Alix-Panabieres C, Pantel K. Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. Cancer Discov 2016;6:479–91.
[11] Jia M, Mao Y, Wu C, Wang S, Zhang H. A platform for primary tumor origin identification of circulating tumor cells via antibody cocktail-based in vivo capture and specific aptamer-based multicolor fluorescence imaging strategy. Anal Chim Acta 2019;1082:136–45.
[12] Haque A, Engel J, Teichmann SA, Lonnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med 2017;9:75.
[13] Lahnemann D, Koster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol 2020;21:31.
[14] Sandberg R. Entering the era of single-cell transcriptomics in biology and medicine. Nat Methods 2014;11:22–4.
[15] Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun 2015;6:8971.
[16] Peng T, Zhu Q, Yin P, Tan K. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. Genome Biol 2019;20:88.
[17] Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput 2017;27:659–78.
[18] Ferreira MM, Ramani VC, Jeffrey SS. Circulating tumor cell technologies. Mol Oncol 2016;10:374–94.
[19] Kowalik A, Kowalewska M, Gozdz S. Current approaches for avoiding the limitations of circulating tumor cells detection methods-implications for diagnosis and treatment of patients with solid tumors. Transl Res 2017;185 (58–84):e15.
[20] Yuan GC, Cai L, Elowitz M, Enver T, Fan G, Guo G, et al. Challenges and emerging directions in single-cell analysis. Genome Biol 2017;18:84.