



Article

Integrative Analysis and Machine Learning based Characterization of Single Circulating Tumor Cells

Arvind Iyer ^{1,†,‡}, Krishan Gupta ^{2,†}, Shreya Sharma ², Kishore Hari ³, Yi Fang Lee ⁴,
Neevan Ramalingam ⁵, Yoon Sim Yap ⁶, Jay West ^{7,§}, Ali Asgar Bhagat ^{4,||,¶},
Balaram Vishnu Subramani ⁸, Burhanuddin Sabuwala ⁹, Tuan Zea Tan ¹⁰, Jean Paul Thiery ¹¹,
Mohit Kumar Jolly ³, Naveen Ramalingam ^{7,*} and Debarka Sengupta ^{1,2,12,*}

- ¹ Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi 110020, India; arvind16122@iiitd.ac.in
- ² Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, New Delhi 110020, India; krishang@iiitd.ac.in (K.G.); shreya15096@iiitd.ac.in (S.S.)
- ³ Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India; kishorehari@iisc.ac.in (K.H.); mkjolly@iisc.ac.in (M.K.J.)
- ⁴ Biolidics Limited, 81 Science Park Drive, 02-03 The Chadwick, Singapore 118257, Singapore; yifang@biolidics.com (Y.F.L.); ali@nus.edu.sg (A.A.B.)
- ⁵ Qualcomm Incorporated, 5775 Morehouse Drive, San Diego, CA 92121, USA; neevan.snap@gmail.com
- ⁶ National Cancer Centre Singapore, 11 Hospital Dr, Singapore 169610, Singapore; Yap.Y.S@nccs.com.sg
- ⁷ Fluidigm Corporation, 2 Tower Place, Suite 2000, South San Francisco, CA 94080, USA; drjayman@gmail.com
- ⁸ School of Mathematics, Indian Institute of Science Education and Research, Thiruvananthapuram 695551, India; balaram.vishnu17@iisertvm.ac.in
- ⁹ Department of Biotechnology, Indian Institute of Technology Madras, Chennai 600036, India; be17b011@smail.iitm.ac.in
- ¹⁰ Cancer Science Institute of Singapore, National University of Singapore, Center for Translational Medicine, Singapore 117599, Singapore; csittz@nus.edu.sg
- ¹¹ Guangzhou Regenerative Medicine and Health; Guangdong laboratory, Chinese Academy of Science, Guangzhou 510530, China; bchtjp@nus.edu.sg
- ¹² Center for Artificial Intelligence, Indraprastha Institute of Information Technology, New Delhi 110020, India
- * Correspondence: naveen.ramalingam@fluidigm.com (N.R.); debarka@iiitd.ac.in (D.S.); Tel.: +91-11-26907446 (D.S.)
- † These authors contributed equally to this work.
- ‡ Current address: Department of Computational Biology, University of Lausanne (UNIL), Lausanne 1015, Switzerland.
- § Current address: BioSkrby Corporation, BioLabs, 701 W Main St, Suite 200, Durham, NC 27701, USA.
- || Current address: Department of Biomedical Engineering, Faculty of Engineering, National University of Singapore, Engineering Drive 1, Singapore 117575, Singapore.
- ¶ Current address: Institute for Health Innovation and Technology (iHealthtech), National University of Singapore, 14 Medical Drive, Singapore 117599, Singapore.

Received: 23 February 2020; Accepted: 16 April 2020; Published: 22 April 2020



Abstract: We collated publicly available single-cell expression profiles of circulating tumor cells (CTCs) and showed that CTCs across cancers lie on a near-perfect continuum of epithelial to mesenchymal (EMT) transition. Integrative analysis of CTC transcriptomes also highlighted the inverse gene expression pattern between PD-L1 and MHC, which is implicated in cancer immunotherapy. We used the CTCs expression profiles in tandem with publicly available peripheral blood mononuclear cell (PBMC) transcriptomes to train a classifier that accurately recognizes CTCs of diverse phenotype. Further, we used this classifier to validate circulating breast tumor cells captured using a newly developed microfluidic system for label-free enrichment of CTCs.

Keywords: high-throughput sequencing; rare cell type; single-cell; RNA-seq; machine learning; CTC; blood

1. Introduction

A staggering 90% of cancer deaths are attributable to metastases [1]. After detaching from solid tumors, cancer cells travel through the bloodstream to reach distant organs and seed the development of metastatic tumors [2]. Cancer cells under circulation are called circulating tumor cells (CTCs) [3]. As a blood-based bio marker, CTCs offer unabated, real-time insights into tumor evolution and therapeutic responses. Despite these promises, the rareness of CTCs in the peripheral blood hinders their isolation and characterization [3]. Cancers in solid tissues develop from epithelial cells, which are typically densely packed in layers. However, dissemination and migration of cancer cells during metastasis require the acquisition of mesenchymal-like features. Transcendence of epithelial cancer cells into mesenchymal-like ones is popularly known as Epithelial to Mesenchymal Transition (EMT).

It is widely understood that due to the loss of epithelial property only a fraction of CTCs can be expected to express canonical epithelial markers such as Epithelial Cell Adhesion Molecule (*EpCAM*). The only FDA (Food and Drug Administration) approved CTC capture platform CELLSEARCH[®] uses epithelial surface marker *EpCAM* to detect CTCs in patients blood [4]. Controlled experiments involving cell-lines have shown that recovery of cells with *EpCAM* expression varies a lot and many canonical epithelial markers are down-regulated in CTCs, undergoing epithelial-mesenchymal transition (EMT) [5]. Therefore, marker-based enrichment techniques are sub-optimal for the comprehensive charting of heterogeneous CTC sub-populations. [6–8] Over the past few years, various CTC capture platforms exploiting biophysical characteristics of cancer cells have been developed [9–11]. *CD45*-based negative enrichment has also been adopted as an alternative strategy. The potential of such antigen-agnostic platforms have not been fully utilized since the chances of immune cell contamination cannot be completely ruled out [9,10]. The recent advent of single-cell RNA sequencing (scRNA-seq) has allowed molecular profiling of single CTCs [12], captured using microfluidic devices [13–17]. Almost all studies that reported molecular profiles of single CTCs resorted to marker based bioinformatic annotation of cell types or applied post-capture staining of CTCs using epithelial/cancer-specific molecular markers [13,18]. To broad base the detection of CTCs, it is therefore important that we develop a scheme to recognize diverse CTC phenotypes presented within a large pool of immune cells.

In this study, we report the ClearCell[®] Polaris[™] workflow that employs size-dependant enrichment of CTCs, followed by negative selection for *CD45* [14,19]. For unbiased labeling of cells of cancer origin, we use publicly available single-cell expression profiles of CTCs and Peripheral Blood Mononuclear Cells (PBMCs) to train a classification system that reliably recognizes a wide variety of CTCs from across different cancer types. In summary, we propose a strategy to employ machine learning based models to detect CTCs retrieved using marker agnostic microfluidic technologies.

2. Materials and Methods

2.1. Description of Datasets

We collected single-cell RNA-seq (scRNA seq) data of circulating tumor cells (CTCs) and peripheral blood mononuclear cells (PBMCs) from 14 different studies in total [2,13,18,20–28] We acquired 558 single CTCs from 10 of these 14 studies. On the other hand, 6 of these studies supplied a total of 37665 PBMCs. Two of these studies with accession numbers GSE67980 and GSE109761 respective offer both blood and CTC transcriptomes. The CTC data entailed five cancer types breast, prostate, melanoma, lung, and pancreas. Notably, circulating breast tumor cells in the data

was supplied by six different studies. Remaining cancer types were represented by single studies (Supplementary Table S1).

2.2. Data Pre-Processing

We downloaded raw read count data for every study from their respective sources (Supplementary Table S1). While merging, we found 15,043 genes common across all the datasets. First, we discarded the poor quality cells that had less than 10% of the genes having non zero expression. The filtering step retained about 5% (1861) of the input cells. Genes with count ≥ 5 in at least 10 cells were retained. A total of 12,335 genes were left after this. Among the 1861 cells, 538 were CTCs. Our final data contained a 12,335 expressed genes and 1861 cells, of which 538 were CTCs. At this stage, we standardized the library depths using median normalization [29–31]. The expression matrix thus obtained was log-transformed after the addition of 1 as pseudo-count. Different gene selection techniques and data used for the various downstream analyses are mentioned in the subsequent sections.

2.3. Construction of Epithelial and Mesenchymal Signatures and E:M Score

While integrating CTC datasets alone, we found 17609 genes common across all 558 CTCs coming from 10 publicly available CTC datasets (Supplementary Table S1). We retained CTCs that expressed at least 5% of the 17609 genes. Genes with read count >5 in at least 10 CTCs were considered for further analyses. At this stage we were left with an expression matrix consisting of 13,600 genes and 554 CTCs. We constructed a panel of 176 well-known epithelial, mesenchymal, and cancer stem cell markers combining information from the CellMarker database [29] and existing literature. The expression matrix of marker genes thus obtained was subjected to stricter criteria for gene and cell selection. We retained 550 cells that expressed at least 10% of these marker genes. Marker genes having minimum read count >5 in at least 30% of these cells were selected for the subsequent analyses. The resulted matrix consisted of 550 cells and 81 marker genes (16 epithelial, 39 mesenchymal, and 26 cancer stem cell markers, see (Supplementary Table S2). We median normalized and log-transformed the generated matrix. For each cell, we computed a comprehensive score for both epithelial and mesenchymal phenotype. To compute the score we first applied Z-score transformation on each cell. To create the signature for specific phenotype, for each cell we combined Z-transformed marker expressions using the below formula.

$$Z_{phenotype} = \frac{\sum_{i \in markers} Z_i}{\sqrt{|markers|}} \quad (1)$$

Here $Z_{phenotype}$ is a comprehensive phenotype specific score computed over individual Z-transformed marker expressions denoted by Z_i , where *markers* denotes the set of markers corresponding to the concerned phenotype. We assigned each single CTC an E:M score by computing the ratio between $Z_{phenotypes}$ computed for epithelial and mesenchymal genes respectively.

2.4. Simulation of E-M Continuum

We identified the regulatory interactions among epithelial (E) and mesenchymal (M) genes under study, together with their connections to canonical regulators of EMT and MET such as the double negative feedback loops involving *miR-200*, *ZEB* and *GRHL2* (Supplementary Note-1). For the constructed network, an ensemble of mathematical models were then created using RACIPE (Random Circuit PErturbation), which considers a set of kinetic parameters randomly chosen from within the biologically relevant ranges [30]. This helps to identify the robust gene expression signatures that can emerge due to given network topology. The simulations were performed in triplets to avoid numerical artifacts/variations due to random sampling. Such an ensemble of models is usually based on ordinary differential equations (ODEs), such as the one mentioned below.

$$\frac{d[VIM]}{dt} = l_{VIM}H^{S+}(ZEB, VIM)H^{S-}(STEP1, VIM) - k_{VIM}[VIM]$$

where $[VIM]$ is the concentration of VIM, and l_{VIM} and k_{VIM} are its production and degradation rates respectively. $H^{S^+}(X, Y)/H^{S^-}(X, Y)$ are the shifted Hill functions that result in up-regulation/down-regulation caused in the expression of Y due to X.

2.5. Classification of Cancer and Blood Transcriptomes

To model the phenotypic identities of CTCs and PBMCs, we trained various classification models. To broad-base our feature selection we used about 3000 cell-type specific markers (Supplementary Table S3) reported in the CellMarker database [29]. Besides, the median normalization we subjected the data to principal component analysis (PCA) [31] and also applied harmony batch correction method [32]. We used three popular classification techniques - Naive Bayes (NB) [33], Gradient Boosting Machines (GBM) [34] and Random Forest (RF) [35] on the training datasets. We evaluated the model on five different datasets: 1. Clearcell-Polaris CTCs; 2. Hydro-Seq Data which uses a novel, hydrodynamic scRNA-seq barcoding technique, for high-throughput CTC capture [11]; 3. the leftover PBMCs, not used for model training; 4. a combination of Clearcell-Polaris and randomly sampled unused 500 PBMC expression profiles; and 5. a combination of Hyrdo-seq data and randomly sampled unused 500 PBMC expression profiles. We computed the accuracy percentage using the equation:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (2)$$

Besides the accuracy percentage, we reported additional model evaluation metrics such as F1 score, Mathews correlation coefficient (MCC) and Cohen's kappa as applicable (Supplementary Table S4).

2.6. Sample Collection

Blood specimens of three *HER2*- (Human epidermal growth factor receptor 2) breast cancer patients (identified as P3, P4, P5) were obtained from the National Cancer Center Singapore, with informed consent following the approved procedures under the institutional review board (IRB) guidelines (CIRB no. 2014/119/B). The clinical sample collection protocols were reviewed and approved by the Sing Health Centralised Institutional Review Board. The determination of estrogen receptor (*ER*), progesterone receptor (*PR*) and human epidermal growth factor receptor 2 (*HER2*) status by immunohistochemistry in this study was based on the latest recommendations of the American Society of Clinical Oncology and the College of American Pathologists. All three subjects had *ER*+/*PR*+/*HER2*- hormone receptor status as analyzed by immunohistochemistry. For P3, blood was drawn (baseline) in August 2016 for CTC enrichment. Following this P3 was on chemotherapy. P4 and P5 were on chemotherapy before their blood samples were collected for CTC enrichment in August and September of 2016, respectively.

2.7. CTC Enrichment

Blood samples were collected in 9 mL of K3EDTA blood collection tubes (Greiner Bio-One, 455036). 6–8.5 mL of whole blood was processed for each run. Red blood cells were first removed with the addition of red blood cell (RBC) lysis buffer (G-Bioscience, St. Louis, MO, USA) and incubation for 10 min at room temperature. Lysed RBCs in the supernatant were discarded after centrifugation. The nucleated cell pellet was suspended in a ClearCell resuspension buffer before CTC enrichment on the ClearCell FX system (Biolidics Limited) [36], performed following manufacturer's instructions.

2.8. Immunofluorescence Suspension Staining

The enriched CTC blood sample was centrifuged at 300 g for 10 min and concentrated to 70 μ L. The cells were stained with the addition of the following markers and antibodies for 1 hour: CellTracker Orange (CTO) (Thermo Fisher, C34551), Calcein AM (Thermo Fisher, L3224), CD45 antibody- conjugated with Alexa 647 (Bio Legend, 304020), and CD31- conjugated with Alexa 647 (Bio

Legend, 303111). 15 μL of RPMI with 10% FBS (Gibco) and 3 μL of RNase inhibitor (Thermo Fisher, N8080119) were also added to improve the viability and RNA quality of the cells. After incubation, 13 mL of PBS was added to dilute the staining reagents. The sample was spun down at 300 g for 10 min and concentrated to 45 μL . In order to achieve optimal buoyancy in an integrated fluidic circuit (IFC), 45 μL of CTCs was mixed with a 30 μL Cell suspension Reagent (Fluidigm, 101-0434) to achieve 75 μL of cell mix.

2.9. Integrated Fluidic Circuit (IFC) Operation

The Polaris IFC is first primed using the Fluidigm Polaris systemTM [19] to fill the control lines on the fluidic circuit, load cell capture beads, and block the inside of PDMS channels to prevent non-specific absorption/adsorption of proteins. To capture and maintain the single cells in the sites, the capture sites (48 sites) are preloaded with beads that are linked on IFC to fabricate a tightly packed bead column during the IFC prime step. After completion of the prime step, the cell mix (cells with suspension reagent) is loaded in three inlets (25 μL each of cell mix) on the Polaris IFC and single cells with CTO+ & Calcein AM+ & CD45- & CD31- are selected to capture sites. Finally, the single cells are processed through template-switching mRNA-seq chemistry for full-length cDNA generation and preamplification on IFC.

2.10. mRNA-Seq Library Preparation and Sequencing

SMARTer[®] Ultra[®] Low RNA Kit for Illumina[®] Sequencing (Clontech[®], 634936) was used to generate preamplified cDNA. The selected and sequestered single cells were lysed using a Polaris cell lysis mixture. The 28- μL cell lysis mix consists of 8.0 μL of Polaris Lysis Reagent (Fluidigm, 101-1637), 9.6 μL of Polaris Lysis Plus Reagent (Fluidigm, 101-1635), 9.0 μL of 3 SMARTTM CDS Primer II A (12 M, Clontech, 634936), and 1.4 μL of Loading Reagent (20X, Fluidigm, 101-1004). The thermal profile for single-cell lysis is 37 $^{\circ}\text{C}$ for 5 min, 72 $^{\circ}\text{C}$ for 3 min, 25 $^{\circ}\text{C}$ for 1 min, and hold at 4 $^{\circ}\text{C}$. The 48- μL preparation volume for reverse transcription (RT) contains 1X SMARTer Kit 5X First-Strand Buffer (5X; Clontech, 634936), 2.5-mM SMARTer Kit Dithiothreitol (100 mM; Clontech, 634936), 1-mM SMARTer Kit dNTP Mix (10 mM each; Clontech, 634936), 1.2- μM SMARTer Kit SMARTer II A Oligonucleotide (12 μM ; Clontech, 634936), 1-U/ μL SMARTer Kit RNase Inhibitor (40 U/ μL ; Clontech, 634936), 10-U/ μL SMARTscribeTM Reverse Transcriptase (100 U/ μL ; Clontech, 634936), and 3.2 μL of Polaris RT Plus Reagent (Fluidigm, 101-1366). All the concentrations correspond to those found in the RT chambers inside the Polaris IFC. The thermal protocol for RT is 42 $^{\circ}\text{C}$ for 90 min (RT), 70 $^{\circ}\text{C}$ for 10 min (enzyme inactivation), and a final hold at 4 $^{\circ}\text{C}$.

The 90- μL preparation volume for PCR contains 1X Advantage 2 PCR Buffer [not short amplicon (SA)](10X, Clontech, 639206, Advantage[®] 2 PCR Kit), 0.4-mM dNTP Mix (50X/10 mM, Clontech, 639206), 0.48- μM IS PCR Primer (12 μM , Clontech, 639206), 2X Advantage 2 Polymerase Mix (50X, Clontech, 639206), and 1X Loading Reagent (20X, Fluidigm, 101-1004). All the concentrations correspond to those found in the PCR chambers inside the Polaris IFC. The thermal protocol for preamplification consists of 95 $^{\circ}\text{C}$ for 1 min (enzyme activation), five cycles (95 $^{\circ}\text{C}$ for 20 s, 58 $^{\circ}\text{C}$ for 4 min, and 68 $^{\circ}\text{C}$ for 6 min), nine cycles (95 $^{\circ}\text{C}$ for 20 s, 64 $^{\circ}\text{C}$ for 30 s, and 68 $^{\circ}\text{C}$ for 6 min), seven cycles (95 $^{\circ}\text{C}$ for 30 s, 64 $^{\circ}\text{C}$ for 30 s, and 68 $^{\circ}\text{C}$ for 7 min), and final extension at 72 $^{\circ}\text{C}$ for 10 min. The preamplified cDNAs are harvested into 48 separate outlets on the Polaris IFC carrier. The cDNA reaction products were then converted into mRNA-seq libraries using the Nextera[®] XT DNA Sample Preparation Kit (Illumina, FC-131-1096 and FC-131-2001, FC-131-2002, FC-131-2003, and FC-131-2004) following the manufacturer's instructions with minor modifications. Specifically, reactions were run at one-quarter of the recommended volume, the tagmentation step was extended to 10 min, and the extension time during the PCR step was increased from 30 to 60 s. After the PCR step, samples were pooled, cleaned twice with 0.9 \times Agencourt AMPure XP SPRI beads (Beckman Coulter), eluted in Tris + EDTA buffer and quantified using a high-sensitivity DNA chip (Agilent). The pooled library was

sequenced on Illumina MiSeq™ using reagent kit v3 (2 × 75 bp paired-end read). The sequencing data generated were processed by standard bioinformatics pipeline (Supplementary Note 2).

2.11. Reference Component Analysis of CTCs and PBMCs

For reference component analysis (RCA), we used the global panels supplied as part of the RCA R package [37]. Each of the global panels consisted of numerous tissue samples. RCA [37] uses cell type specific genes for measuring the correlation between the tissue types and the input single cells. Due to the low amount of starting RNA, single cell expression data is far noisier than bulk expression data. As a result, tissue types represented by lowly expressed feature genes can potentially give rise to significant levels of noise. In each global panel, we, therefore, retained 50% of the tissue types with the highest median expression of the feature genes. RCA [37] analysis provided us with both single cell-tissue correlation heat-map and 2D projection of the individual transcriptomes.

2.12. Data and Code Availability

The data-set used in the study are available from links mentioned in the (Supplementary Table S1). Single cell sequencing data generated for this paper is deposited at GEO with accession number [GSE129474](#). Code used for analysis is available at this [link](#) and a R package is available at [link](#).

3. Results

3.1. Integration of Single Cell Expression Datasets of Circulating Tumor Cells

We collected about 500 single CTC transcriptomes from 10 independent studies, representing five different cancer types i.e., breast, prostate, lung, pancreas, and melanoma (Figure 1B, Supplementary Table S1). On the other hands, as control, expression profiles of human PBMCs were collected from six different studies (Supplementary Table S1). About 70% of the CTCs came from various breast cancer studies. CTC datasets that we curated were of variable quality. We preprocessed the data to ensure that the poor-quality cells and unexpressed genes were discarded (Methods, Supplementary Figure S1). We further normalised the combined expression matrix to control for the library depth (Methods). We tracked expression of some of the canonical epithelial (*KRT8*, *KRT18*, *EpCAM*, *CDH1*) and leukocyte markers (*PTPRC*, *VIM*) to cross-validate the cell type identities. Elevated expression levels of a subset of epithelial markers were observed in a vast majority of the CTCs (Figure 1C, Supplementary Figure S2). Significant up-regulation of platelet and fibroblast markers was observed in large fractions of CTCs (Figure 1C, Supplementary Figure S2). This combined data source served as the basis for the majority of our analysis and development of the CTC-immune cell classification system (Figure 1A).

3.2. Ubiquity of Epithelial-Mesenchymal Transition in Cancer Metastasis

Epithelial-mesenchymal transition (EMT) and mesenchymal-epithelial transition (MET) have long been postulated to play key roles in cancer metastasis and drug resistance [38]. The integration of CTC datasets presented us with the opportunity to probe into its validity. For each CTC, we computed two scores indicating the strength of epithelial and mesenchymal phenotypes respectively (Methods). In this analysis, we used tens of canonical markers of each of the concerned phenotypes. We detected near-perfect anti-correlation of ($\rho = -0.91$) the phenotypes across CTCs, coming from all cancer types (Figure 2A, Supplementary Figure S3). Our findings were consistent when we tracked the association between these phenotypes for CTCs from individual studies (Supplementary Figure S4). Notably, CTC transcriptomes were frequently found on a continuum of epithelial-mesenchymal transition in most of the datasets (Figure 2B). However, a agglomerative hierarchical clustering stratified the CTCs into two groups largely based on their approximate binarized identity as epithelial/mesenchymal cells (Supplementary Figure S13). In selected studies, in spite of being on a continuum, CTCs were found to form clusters towards the epithelial and the mesenchymal poles respectively (Supplementary

Figure S4). Melanocytes derive from a highly invasive, multipotent embryonic cell population called the neural crest. It is suggested that the high degree of plasticity and the aggressiveness of malignant melanoma originate due to the re-activation of the embryonic neural crest program, which is silenced in due course of normal melanocyte differentiation [39].

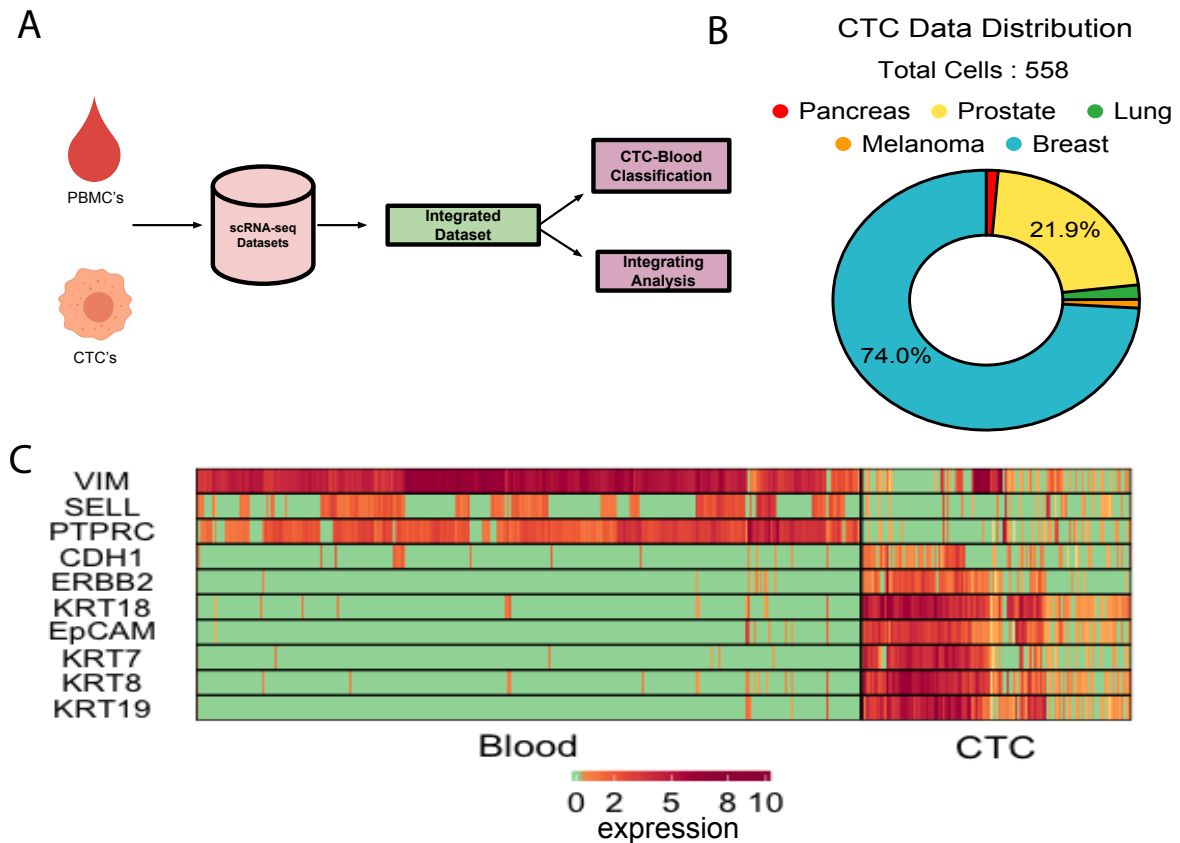


Figure 1. Integrative analysis of CTC transcriptomes: (A) Schematic of study. (B) Cancer types represented by the integrated CTC population. (C) Expression of canonical epithelial and immune cell markers in CTCs and the PBMCs under study.

Unlike the CTCs of most cancer types, circulating melanoma cells were found to be clustered exclusively around the mesenchymal pole of the E-M continuum (Supplementary Figure S4). Our E:M scores were found to be correlated (negatively) ($\rho = -0.779$) with EMT score as proposed by Tan and colleagues [40] (Figure 2C). One should note that a CTC, enriched with epithelial markers would receive a large positive E:M score, and a large negative EMT score. As a secondary validation, we constructed a network incorporating regulations among E and M genes under study (Methods, Supplementary Figure S5). Simulation experiments on this network using Ordinary Differential Equations (ODE) resulted in expression anti-correlation ($\rho = -0.65$) between *CDH1* and *VIM* (Methods, Figure 2D, Supplementary Figure S6).

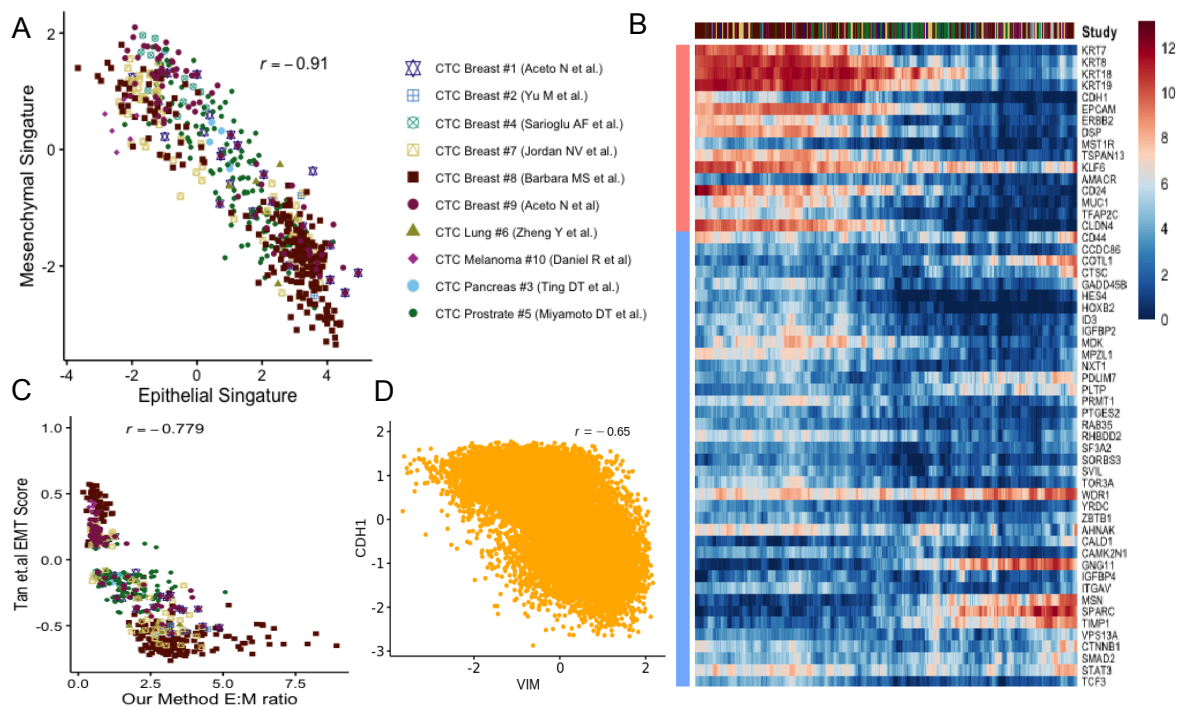


Figure 2. Epithelial-mesenchymal transition in cancer metastasis: (A) Scatter plot showing anti-correlation between epithelial and mesenchymal phenotypes across studies. (B) The moving average smoothen log(expression+1) of CTC dataset on epithelial and mesenchymal markers where cells are ordered based on their respective E:M score as described in the main methods. (C) Scatter diagram depicting the correspondence between E:M score and the EMT score proposed by Tan and colleagues [40]. (D) CDH1-VIM anti-correlation observed due to simulation of EMT associated regulatory network.

3.3. Clear Patterns Observed in Expression Gradient of Immune Check-Point Inhibitor and Stemness Marker

The activation of HLA class I (HLA-I) antigens on tumor cells is essential for the activation of cytotoxic T-lymphocytes. It has been demonstrated in mouse lines as well as human cancers that during natural cancer progression tumors gradually lose MHC-I expression as a result of a T-cell mediated immune selection [41]. On the other hand, the PD-1/PD-L1 pathway represents an adaptive immune resistance mechanism exerted by tumor cells in response to endogenous immune anti-tumor activity. PD-L1 expressed by tumor cells binds to PD-1 receptors on the activated T cells, which leads to the inhibition of the cytotoxic T cells [42]. Taken together, the loss of major histocompatibility complex (MHC) proteins (aka HLAs) and the activation of PD-L1 signify the prevention of cytotoxic T cell activities on tumor cells. Of late, immune checkpoint inhibitors, targeting the PD-1/PD-L1 pathway, have emerged as successful cancer treatment options [43]. In our curated datasets, we found only a minor fraction of CTCs expressing PD-L1. However, PD-L1-MHC anti correlation was evident across studies (Figure 3A). One of the datasets containing the maximum number of PD-L1-activated breast CTCs showed concurrence of PD-L1 with mesenchymal phenotype (Supplementary Figure S7). To date, multiple studies have linked EMT to the formation of cancer stem cells (CSCs). In a seminal paper, Mani and colleagues demonstrated the generation of a $CD44^{high}/CD24^{low}$, mammary stem cell-like population due to the induction of EMT. These cells were able to initiate tumors quite efficiently in the mouse. We tracked expression changes in CSC markers along E-M continuum [44]. $CD44^{high}/CD24^{low}$ CTCs indeed emerge late in the spectrum, following EMT induction (Figure 3b). This demonstrates how integrative analysis of CTC transcriptomes can help pinpoint stem-like phenotypes, with high tumorigenesis potential.

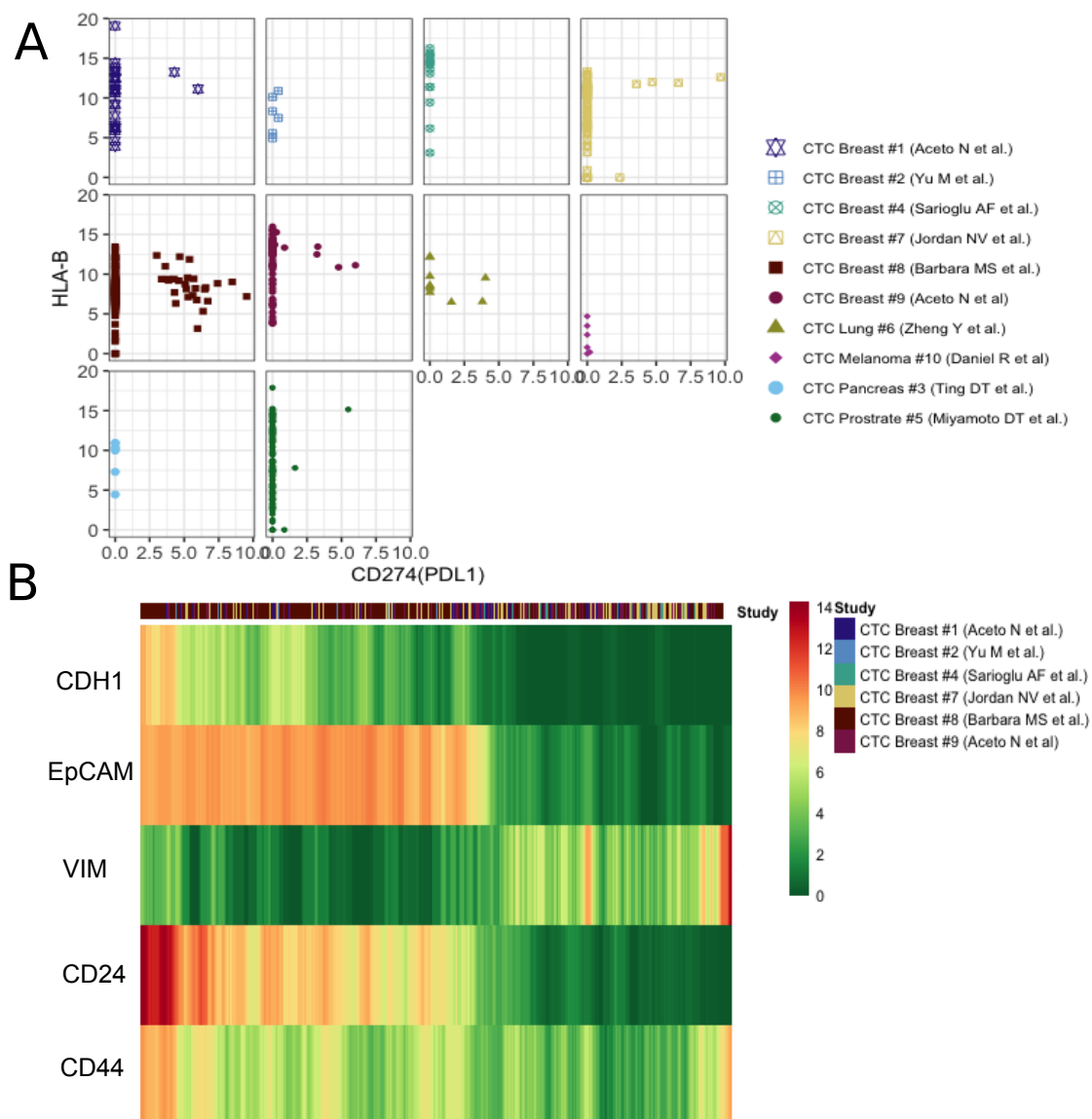


Figure 3. Patterns observed in expression gradient of immune check-point inhibitor and stemness markers. **(A)** The scatter plot of PDL1 and HLA-B expression in each study. **(B)** The moving average smoothen $\log(\text{expression}+1)$ of well known specific epithelial (CDH1, EpCAM), mesenchymal (VIM) and cancer stem cell markers (CD24, CD44) across breast CTCs, ordered based on the ratio of epithelial and mesenchymal signatures calculated as described in the main methods.

3.4. CTC-PBMC Classification System

We trained a classifier on publicly available single cell expression profiles of human CTCs and PBMCs. Expression datasets curated from independent studies were subjected to rigorous data preprocessing steps (Methods). Notably, the state of the art batch effect removal method harmony [32] failed to improve the performance of the classification algorithms, compared to a simple median normalisation baseline (Supplementary Figure S12). We compared the performance of three classifiers—Naïve Bayes [33], Random Forest [35], and Gradient Boosting Machine [34]. We evaluated the model on five different datasets (Methods). Overall, the best performing model was GBM with a mean accuracy of $\sim 93\%$ (Figure 4B). Notably, expression profiles of the CTCs retrieved by the Clearcell-Polaris system were all predicted as CTCs. $\sim 80\%$ CTCs captured by the recently developed Hydro-Seq [11] (a hydrodynamic RNA-seq barcoding technique, for high-throughput CTC analysis) technique were classified as CTCs (Supplementary Table S4).

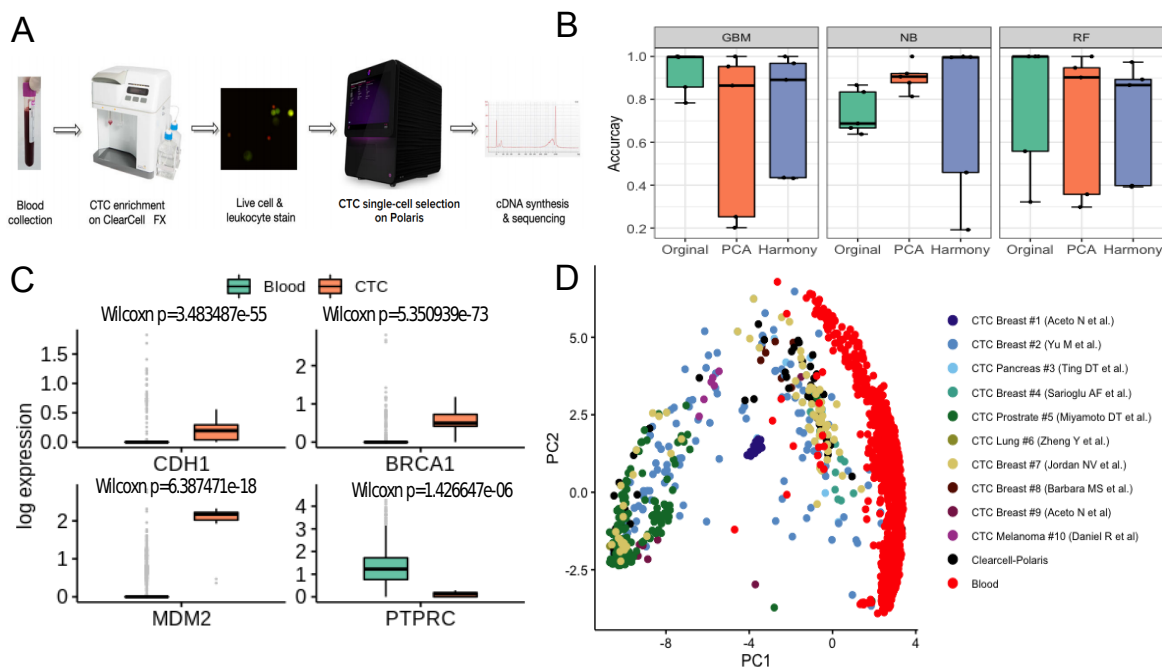


Figure 4. Label-free detection and characterisation of CTCs. **(A)** ClearCell-Polaris workflow involving size-based CTC enrichment by ClearCell FX system, followed by single cell selection and CD45/CD31 depletion using Polaris. **(B)** Performance of various machine learning algorithms in distinguishing between CTCs and PBMCs. Cells in each dataset were tested against a classifier trained on the remaining datasets. Box plots show the prediction accuracy's for different choices of classification algorithms (Naive Bayes or NB, Random Forest or RF, Gradient Boosting Machine or GBM) and normalisation/batch-effect correction methods. **(C)** Box-plots showing canonical epithelial/breast cancer specific markers, up-regulated in the CTC population compared to the PBMCs. As expected, PTPRC, a pan leukocyte maker shows elevated expression levels in PBMCs as compared to CTCs. **(D)** Reference Component Analysis (RCA) based 2D projection of CTCs. PBMCs (red) are visibly separated from CTCs. CTCs enriched using the ClearCell-Polaris workflow cluster with CTCs of other types.

3.5. Identification of CTCs Captured Using Novel Label-Free Microfluidic Workflow

Existing technologies enrich CTCs with some level of contaminating white blood cells (WBCs). This poses a significant challenge in differentiating CTCs from immune cells. We addressed this challenge by integrating two commercially available microfluidic systems namely Biolidics ClearCell FX System [36] and the Fluidigm Polaris™ system [19] (Methods, Figure 4A). In the proposed workflow CTCs are enriched in two steps - size-based enrichment by ClearCell, followed by CD45 (leukocyte marker) and CD31 (endothelial cell marker) based negative selection by Polaris [19].

To validate the workflow and the accompanying PBMC-CTC classification system, we processed peripheral blood samples of three HER2-, stage IV breast cancer patients (identified as P3, P4, P5) through the microfluidic device ensemble (Methods, Supplementary Figure S8). Polaris could retrieve 13, 12 and 32 cells from the blood samples of patients P3, P4, P5 respectively. 15 of these 57 cells passed the filtering criteria (Supplementary Figure S9). All 15 cells were classified as CTCs. We used additional validation criteria to determine the carcinogenic origin of the captured cells. When compared to a set of randomly selected PBMCs, ClearCell Polaris captured cells showed elevated expression of breast cancer-specific markers *BRCA1* and *MDM2* (p -value < 0.05) [45] (Figure 4C). We also detected up-regulation of *CDH1*, a canonical epithelial cell marker. Expression of *CD45* (*PTPRC*) was considerably low in these cells compared to the PBMC transcriptomes (p -value < 0.05) (Figure 4C). Reference component analysis (RCA) allows noise-free single cell clustering, by projecting single cell transcriptomes on reference bulk expression data. We subjected all CTC and PBMC transcriptomes to

RCA analysis [37]. ClearCell-Polaris captured CTCs grouped with other CTCs, whereas the PBMCs formed a separate cluster (Methods, Figure 4d, Supplementary Figure S10).

4. Discussion

CTCs have been shown to be of prognostic significance in patients with various cancers [2,18,28]. We integrated single-cell expression profiles from various published studies and analyzed the emergence of epithelial to mesenchymal transition among CTCs. For this, we developed the E:M score that ordered CTC transcriptomes on an approximate pseudo-temporal axis of epithelial-mesenchymal transition. Our proposed EMT scoring method, in principle, is similar to the method proposed by Tan and colleagues, which focuses on six major cancer types, namely ovarian, breast, bladder, colorectal, gastric, and lung. Different from this, we used widely accepted, literature curated E and M markers agnostic of the cancer types. Although both the methods correlate well when applied to the CTC transcriptomes (Figure 2C), we found our proposed methods depict the E to M continuum better (Figure 2B and Supplementary Figure S14).

It is suspected that a large number of CTCs do not portray the signature of cancer epithelium, largely due to their acquired phenotype that is suitable for migration [28]. We leveraged the power of machine learning in techniques in reliably distinguishing CTCs from other relatively way more abundant immune cell types. This is achieved by the integration of publicly available CTC datasets and machine learning-based model training. We provide a user-friendly R package for CTC classification that provides a probabilistic score indicating the cancer origin of individual cells. Our reported ClearCell® Polaris™ workflow, in tandem with the machine learning based CTC-immune cell classification system, for the first time, enables truly unbiased detection of circulating tumor cells. With declining per cell cost associated with single-cell gene expression screening, we speculate a high adoption rate for our proposed strategy.

An integrative study of CTC transcriptomes presented us with the opportunity to discover consistent pan-cancer CTC surface-proteins, besides *EpCAM*. We looked for surface-protein coding genes that are differentially upregulated in CTCs over blood cells (Supplementary Note-3). Most remarkable among these were *ITGB5*, *TACSTD2*, *SLC39A6* (Supplementary Figure S12). In addition to *EpCAM*, some of these markers might be useful to broad-base marker dependent capture of CTCs.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2077-0383/9/4/1206/s1>, Supplementary Note 1: Network analysis to investigate the mechanistic basis of EMT continuum phenotype observed in the data analysis, Supplementary Note 2: Gene expression quantification of CTCs detected by the ClearCell Polaris workflow, Supplementary Note 3: Exploration of novel surface markers for CTCs, Supplementary Figure S1: Data Quality of studies, Supplementary Figure S2: Expression of known markers in curated CTCs and PBMCs, Supplementary Figure S3: Combined epithelial, mesenchymal and cancer stem cell signatures, Supplementary Figure S4: Scatter plots show Epithelial-Mesenchymal anti-correlation for individual datasets, Supplementary Figure S5: The network simulated using RACIPE, Supplementary Figure S6: Random network simulation results, Supplementary Figure S7: Expression Gradient of Immune Check-Point Inhibitor and Stemness Marker, Supplementary Figure S8: Treatment history of the patients, Supplementary Figure S9: Number of expressed genes in CTCs detected using the Clearcell-Polaris workflow. Supplementary Figure S10: Tissue - single cell correlation plot obtained from RCA, Supplementary Figure S11: Log2 fold change of surface markers between CTC and PBMC populations, Supplementary Figure S12: PCA plots of log transformed median normalized counts and Harmony batch correction method, Supplementary Figure S13: Clustered heatmap of Main Figure 2B, Supplementary Figure S14: Continuum plot using Tan et al method, Supplementary Table S1: List of all studies from which datasets are used, Supplementary Table S2: Functional details of the EMT related genes used in the study, Supplementary Table S3: Genes used as features for machine learning based analyses, Supplementary Table S4: Machine learning results.

Author Contributions: D.S. and N.R. (Naveen Ramalingam) conceived the project. A.I. and K.G. performed the majority of the analyses under the supervision of D.S. S.S. assisted A.I. and K.G. in the computational analyses. T.Z.T. and J.P.T. conceived and computed the EMT scores. M.K.J. planned the EMT modeling. K.H., B.S., B.V.S. performed the associated analysis under M.K.J.'s supervision. N.R. (Naveen Ramalingam), J.W., A.A.B. conceived integration of FX and Polaris. N.R. (Naveen Ramalingam) and Y.F.L. developed the label-free workflow. Y.S.Y. provided the patient samples. Y.F.L. tested patient samples and N.R. (Naveen Ramalingam) assisted N.R. (Naveen Ramalingam) in data analysis. All the authors discussed the results, co-wrote and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the INSPIRE Faculty Grant (DST/INSPIRE/04/2015/003068) awarded to D.S. by the Department of Science and Technology (DST), Govt. of India. M.K.J is supported by Ramanujan Fellowship provided by SERB, DST, Government of India (SB/S2/RJN-049/2018).

Conflicts of Interest: NR is an employee and stockholder of Fluidigm Corporation. AAB and YFL are employees of Biolidics Ltd and are stockholders in the company. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Seyfried, T.N.; Huysentruyt, L.C. On the origin of cancer metastasis. *Crit. Rev. Oncog.* **2013**, *18*, 43. [[CrossRef](#)] [[PubMed](#)]
2. Song, Y.; Tian, T.; Shi, Y.; Liu, W.; Zou, Y.; Khajvand, T.; Wang, S.; Zhu, Z.; Yang, C. Enrichment and single-cell analysis of circulating tumor cells. *Chem. Sci.* **2017**, *8*, 1736–1751. [[CrossRef](#)]
3. Dive, C.; Brady, G. SnapShot: Circulating tumor cells. *Cell* **2017**, *168*, 742–742. [[CrossRef](#)] [[PubMed](#)]
4. Andreopoulou, E.; Yang, L.Y.; Rangel, K.; Reuben, J.; Hsu, L.; Krishnamurthy, S.; Valero, V.; Fritsche, H.; Cristofanilli, M. Comparison of assay methods for detection of circulating tumor cells in metastatic breast cancer: AdnaGen AdnaTest BreastCancer Select/Detect™ versus Veridex CellSearch™ system. *Int. J. Cancer* **2012**, *130*, 1590–1597. [[CrossRef](#)] [[PubMed](#)]
5. Mikolajczyk, S.D.; Millar, L.S.; Tsinberg, P.; Coutts, S.M.; Zomorodi, M.; Pham, T.; Bischoff, F.Z.; Pircher, T.J. Detection of EpCAM-negative and cytokeratin-negative circulating tumor cells in peripheral blood. *J. Oncol.* **2011**, *2011*, 252361. [[CrossRef](#)]
6. Miller, M.C.; Doyle, G.V.; Terstappen, L.W. Significance of circulating tumor cells detected by the CellSearch system in patients with metastatic breast colorectal and prostate cancer. *J. Oncol.* **2010**, *2010*, 617421. [[CrossRef](#)]
7. Farace, F.; Massard, C.; Vimond, N.; Drusch, F.; Jacques, N.; Billiot, F.; Laplanche, A.; Chauchereau, A.; Lacroix, L.; Planchard, D.; et al. A direct comparison of CellSearch and ISET for circulating tumour-cell detection in patients with metastatic carcinomas. *Br. J. Cancer* **2011**, *105*, 847–853. [[CrossRef](#)]
8. Wang, L.; Balasubramanian, P.; Chen, A.P.; Kummar, S.; Evrard, Y.A.; Kinders, R.J. Promise and limits of the CellSearch platform for evaluating pharmacodynamics in circulating tumor cells. *Semin. Oncol.* **2016**, *43*, 464–475. [[CrossRef](#)]
9. Gabriel, M.T.; Calleja, L.R.; Chalopin, A.; Ory, B.; Heymann, D. Circulating tumor cells: A review of non-EpCAM-based approaches for cell enrichment and isolation. *Clin. Chem.* **2016**, *62*, 571–581. [[CrossRef](#)]
10. Ferreira, M.M.; Ramani, V.C.; Jeffrey, S.S. Circulating tumor cell technologies. *Mol. Oncol.* **2016**, *10*, 374–394. [[CrossRef](#)]
11. Cheng, Y.H.; Chen, Y.C.; Lin, E.; Brien, R.; Jung, S.; Chen, Y.T.; Lee, W.; Hao, Z.; Sahoo, S.; Kang, H.M.; et al. Hydro-Seq enables contamination-free high-throughput single-cell RNA-sequencing for circulating tumor cells. *Nat. Commun.* **2019**, *10*, 2163. [[CrossRef](#)] [[PubMed](#)]
12. Chen, X.X.; Bai, F. Single-cell analyses of circulating tumor cells. *Cancer Biol. Med.* **2015**, *12*, 184.
13. Sarioglu, A.F.; Aceto, N.; Kojic, N.; Donaldson, M.C.; Zeinali, M.; Hamza, B.; Engstrom, A.; Zhu, H.; Sundaresan, T.K.; Miyamoto, D.T.; et al. A microfluidic device for label-free, physical capture of circulating tumor cell clusters. *Nat. Methods* **2015**, *12*, 685. [[CrossRef](#)] [[PubMed](#)]
14. Warkiani, M.E.; Guan, G.; Luan, K.B.; Lee, W.C.; Bhagat, A.A.S.; Chaudhuri, P.K.; Tan, D.S.W.; Lim, W.T.; Lee, S.C.; Chen, P.C.; et al. Slanted spiral microfluidics for the ultra-fast, label-free isolation of circulating tumor cells. *Lab A Chip* **2014**, *14*, 128–137. [[CrossRef](#)] [[PubMed](#)]
15. Karabacak, N.M.; Spuhler, P.S.; Fachin, F.; Lim, E.J.; Pai, V.; Ozkumur, E.; Martel, J.M.; Kojic, N.; Smith, K.; Chen, P.i.; et al. Microfluidic, marker-free isolation of circulating tumor cells from blood samples. *Nat. Protoc.* **2014**, *9*, 694. [[CrossRef](#)] [[PubMed](#)]
16. Xu, L.; Mao, X.; Imrali, A.; Syed, F.; Mutsvangwa, K.; Berney, D.; Cathcart, P.; Hines, J.; Shamash, J.; Lu, Y.J. Optimization and evaluation of a novel size based circulating tumor cell isolation system. *PLoS ONE* **2015**, *10*, e0138032. [[CrossRef](#)] [[PubMed](#)]
17. Warkiani, M.E.; Khoo, B.L.; Wu, L.; Tay, A.K.P.; Bhagat, A.A.S.; Han, J.; Lim, C.T. Ultra-fast, label-free isolation of circulating tumor cells from blood using spiral microfluidics. *Nat. Protoc.* **2016**, *11*, 134. [[CrossRef](#)]

18. Aceto, N.; Bardia, A.; Miyamoto, D.T.; Donaldson, M.C.; Wittner, B.S.; Spencer, J.A.; Yu, M.; Pely, A.; Engstrom, A.; Zhu, H.; et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **2014**, *158*, 1110–1122. [[CrossRef](#)]
19. Ramalingam, N.; Fowler, B.; Szpankowski, L.; Leyrat, A.A.; Hukari, K.; Maung, M.T.; Yorza, W.; Norris, M.; Cesar, C.; Shuga, J.; et al. Fluidic logic used in a systems approach to enable integrated single-cell functional analysis. *Front. Bioeng. Biotechnol.* **2017**, *4*, 70. [[CrossRef](#)]
20. Lin, E.; Cao, T.; Nagrath, S.; King, M.R. Circulating tumor cells: Diagnostic and therapeutic applications. *Annu. Rev. Biomed. Eng.* **2018**, *20*, 329–352. [[CrossRef](#)]
21. Aceto, N.; Bardia, A.; Wittner, B.S.; Donaldson, M.C.; O’Keefe, R.; Engstrom, A.; Bersani, F.; Zheng, Y.; Comaills, V.; Niederhoffer, K.; et al. AR expression in breast cancer CTCs associates with bone metastases. *Mol. Cancer Res.* **2018**, *16*, 720–727. [[CrossRef](#)] [[PubMed](#)]
22. Zheng, Y.; Miyamoto, D.T.; Wittner, B.S.; Sullivan, J.P.; Aceto, N.; Jordan, N.V.; Yu, M.; Karabacak, N.M.; Comaills, V.; Morris, R.; et al. Expression of β -globin by cancer cells promotes cell survival during blood-borne dissemination. *Nat. Commun.* **2017**, *8*, 14344. [[CrossRef](#)] [[PubMed](#)]
23. Ting, D.T.; Wittner, B.S.; Ligorio, M.; Jordan, N.V.; Shah, A.M.; Miyamoto, D.T.; Aceto, N.; Bersani, F.; Brannigan, B.W.; Xega, K.; et al. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* **2014**, *8*, 1905–1918. [[CrossRef](#)] [[PubMed](#)]
24. Miyamoto, D.T.; Zheng, Y.; Wittner, B.S.; Lee, R.J.; Zhu, H.; Broderick, K.T.; Desai, R.; Fox, D.B.; Brannigan, B.W.; Trautwein, J.; et al. RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science* **2015**, *349*, 1351–1356. [[CrossRef](#)]
25. Van der Wijst, M.G.; Brugge, H.; de Vries, D.H.; Deelen, P.; Swertz, M.A.; Franke, L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **2018**, *50*, 493–497. [[CrossRef](#)]
26. Jordan, N.V.; Bardia, A.; Wittner, B.S.; Benes, C.; Ligorio, M.; Zheng, Y.; Yu, M.; Sundaresan, T.K.; Licausi, J.A.; Desai, R.; et al. HER2 expression identifies dynamic functional states within circulating breast cancer cells. *Nature* **2016**, *537*, 102–106. [[CrossRef](#)]
27. Gkoutela, S.; Castro-Giner, F.; Szczerba, B.M.; Vetter, M.; Landin, J.; Scherrer, R.; Krol, I.; Scheidmann, M.C.; Beisel, C.; Stirnimann, C.U.; et al. Circulating Tumor Cell Clustering Shapes DNA Methylation to Enable Metastasis Seeding. *Cell* **2019**, *176*, 98–112. [[CrossRef](#)]
28. Szczerba, B.M.; Castro-Giner, F.; Vetter, M.; Krol, I.; Gkoutela, S.; Landin, J.; Scheidmann, M.C.; Donato, C.; Scherrer, R.; Singer, J.; et al. Neutrophils escort circulating tumour cells to enable cell cycle progression. *Nature* **2019**, *566*, 553–557. [[CrossRef](#)] [[PubMed](#)]
29. Zhang, X.; Lan, Y.; Xu, J.; Quan, F.; Zhao, E.; Deng, C.; Luo, T.; Xu, L.; Liao, G.; Yan, M.; et al. CellMarker: A manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **2018**, *47*, D721–D728. [[CrossRef](#)]
30. Huang, B.; Jia, D.; Feng, J.; Levine, H.; Onuchic, J.N.; Lu, M. RACIPE: A computational tool for modeling gene regulatory circuits using randomization. *BMC Syst. Biol.* **2018**, *12*, 74. [[CrossRef](#)]
31. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *London Edinburgh Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [[CrossRef](#)]
32. Korsunsky, I.; Millard, N.; Fan, J.; Slowikowski, K.; Zhang, F.; Wei, K.; Baglaenko, Y.; Brenner, M.; Loh, P.R.; Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **2019**, *16*, 1289–1296. [[CrossRef](#)] [[PubMed](#)]
33. Rish, I. An empirical study of the naive Bayes classifier. In Proceedings of the IJCAI 2001 workshop on empirical methods in artificial intelligence, Seattle, DC, USA, 4 August 2001; Volume 3, pp. 41–46.
34. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Annals Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
35. Ho, T.K. Random decision forests. In Proceedings of the 3rd international conference on document analysis and recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282.
36. Lee, Y.; Guan, G.; Bhagat, A.A. ClearCell[®] FX, a label-free microfluidics technology for enrichment of viable circulating tumor cells. *Cytom. Part A* **2018**, *93*, 1251–1254. [[CrossRef](#)] [[PubMed](#)]
37. Li, H.; Courtois, E.T.; Sengupta, D.; Tan, Y.; Chen, K.H.; Goh, J.J.L.; Kong, S.L.; Chua, C.; Hon, L.K.; Tan, W.S.; et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **2017**, *49*, 708. [[CrossRef](#)]

38. Nieto, M.A.; Huang, R.Y.J.; Jackson, R.A.; Thiery, J.P. EMT: 2016. *Cell* **2016**, *166*, 21–45. [[CrossRef](#)]
39. Bailey, C.M.; Morrison, J.A.; Kulesa, P.M. Melanoma revives an embryonic migration program to promote plasticity and invasion. *Pigment Cell Melanoma Res.* **2012**, *25*, 573–583. [[CrossRef](#)]
40. Tan, T.Z.; Miow, Q.H.; Miki, Y.; Noda, T.; Mori, S.; Huang, R.Y.J.; Thiery, J.P. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* **2014**, *6*, 1279–1293. [[CrossRef](#)]
41. Garrido, F.; Ruiz-Cabello, F.; Aptsiauri, N. Rejection versus escape: The tumor MHC dilemma. *Cancer Immunol. Immunother.* **2017**, *66*, 259–271. [[CrossRef](#)]
42. Pardoll, D.M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **2012**, *12*, 252–264. [[CrossRef](#)]
43. Gong, J.; Chehrazhi-Raffle, A.; Reddi, S.; Salgia, R. Development of PD-1 and PD-L1 inhibitors as a form of cancer immunotherapy: A comprehensive review of registration trials and future considerations. *J. Immunother. Cancer* **2018**, *6*, 8. [[CrossRef](#)] [[PubMed](#)]
44. Mani, S.A.; Guo, W.; Liao, M.J.; Eaton, E.N.; Ayyanan, A.; Zhou, A.Y.; Brooks, M.; Reinhard, F.; Zhang, C.C.; Shipitsin, M.; et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **2008**, *133*, 704–715. [[CrossRef](#)] [[PubMed](#)]
45. Parker, J.S.; Mullins, M.; Cheang, M.C.; Leung, S.; Voduc, D.; Vickery, T.; Davies, S.; Fauron, C.; He, X.; Hu, Z.; et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **2009**, *27*, 1160. [[CrossRef](#)] [[PubMed](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).