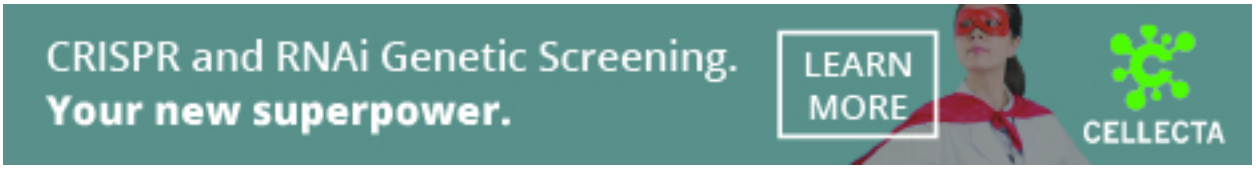
Marker-free characterization of full-length transcriptomes of single live circulating tumor cells

Sarita Poonia, Anurag Goel, Smriti Chawla, et al.

Genome Res. 2023 33: 80-95 originally published online November 22, 2022
Access the most recent version at doi:10.1101/gr.276600.122



Method

Marker-free characterization of full-length transcriptomes of single live circulating tumor cells

Sarita Poonia,[1] Anurag Goel,[2,3] Smriti Chawla,[1] Namrata Bhattacharya,[2] Priyadarshini Rai,[1] Yi Fang Lee,[4,11] Yoon Sim Yap,[5] Jay West,[6,12] Ali Asgar Bhagat,[4,13,14] Juhi Tayal,[7] Anurag Mehta,[8] Gaurav Ahuja,[1] Angshul Majumdar,[2,9,10] Naveen Ramalingam,[6] and Debarka Sengupta[1,2,9]

[1]*Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi 110020, India;* [2]*Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi 110020, India;* [3]*Department of Computer Science and Engineering, Delhi Technological University, New Delhi 110042, India;* [4]*Biolidics Limited, Singapore 118257, Singapore;* [5]*National Cancer Centre Singapore, Singapore 169610, Singapore;* [6]*Fluidigm Corporation, South San Francisco, California 94080, USA;* [7]*Department of Research,* [8]*Department of Laboratory Services and Molecular Diagnostics, Rajiv Gandhi Cancer Institute and Research Centre-Delhi (RGCIRC-Delhi), New Delhi 110085, India;* [9]*Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi 110020, India;* [10]*Department of Electronics & Communications Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi 110020, India*

The identification and characterization of circulating tumor cells (CTCs) are important for gaining insights into the biology of metastatic cancers, monitoring disease progression, and medical management of the disease. The limiting factor in the enrichment of purified CTC populations is their sparse availability, heterogeneity, and altered phenotypes relative to the primary tumor. Intensive research both at the technical and molecular fronts led to the development of assays that ease CTC detection and identification from peripheral blood. Most CTC detection methods based on single-cell RNA sequencing (scRNA-seq) use a mix of size selection, marker-based white blood cell (WBC) depletion, and antibodies targeting tumor-associated antigens. However, the majority of these methods either miss out on atypical CTCs or suffer from WBC contamination. We present unCTC, an R package for unbiased identification and characterization of CTCs from single-cell transcriptomic data. unCTC features many standard and novel computational and statistical modules for various analyses. These include a novel method of scRNA-seq clustering, named deep dictionary learning using *k*-means clustering cost (DDLK), expression-based copy number variation (CNV) inference, and combinatorial, marker-based verification of the malignant phenotypes. DDLK enables robust segregation of CTCs and WBCs in the pathway space, as opposed to the gene expression space. We validated the utility of unCTC on scRNA-seq profiles of breast CTCs from six patients, captured and profiled using an integrated ClearCell FX and Polaris workflow that works by the principles of size-based separation of CTCs and marker-based WBC depletion.

[Supplemental material is available for this article.]

Cancer ranks as a prime reason for death and a vital barrier to longer life expectancy in every country of the world (Sung et al. 2021). According to World Health Organization (WHO) estimates, in 2019 (Mathers 2020) among 183 countries, cancer ranked as the first or second cause of death of people below the age of 70 yr and ranked third or fourth in 23 countries (Sung et al. 2021). The primary reason for 90% of cancer-related deaths is metastasis (Bittner et al. 2020), the process in which the cancer cells detach from the primary tumor, enter into the circulation, and eventually colonize distant organs, causing the spread of disease (Krebs et al. 2014; Siegel et al. 2015). To metastasize, cancer cells secrete chemokines to attract immune cells (McAllister and Weinberg 2014; Liu and Cao 2016), facilitating tumor proliferation and intravasation (Gajewski et al. 2013; Kitamura 2018). After cancer cells enter the bloodstream, they are subjected to various stressors, including the lack of cell–cell and cell–matrix adhesion, shear pressures, and immune response. Despite this, a few cancer cells make it through the tortuous journey and leave the vasculature to a secondary site (Shenoy and Lu 2016; Follain et al. 2018).

Circulating tumor cells (CTCs) have recently attracted a lot of attention owing to their critical role in tumor metastasis. Around 40% to 80% of patients with metastatic breast cancers have been found to have CTCs in their blood (Kwa and Esteva 2018). The

Present addresses: [11]Thermo Fisher Scientific, Singapore 739256, Singapore; [12]BioSkryb Corporation, Durham, NC 27701, USA; [13]Department of Biomedical Engineering, Faculty of Engineering, National University of Singapore, Singapore 117575, Singapore; [14]Institute for Health Innovation and Technology (iHealthtech), National University of Singapore, Singapore 117599, Singapore
Corresponding authors: debarka@iiitd.ac.in, naveen.ramalingam@fluidigm.com, angshul@iiitd.ac.in
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276600.122.

33:80–95 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/23; www.genome.org

detection and characterization of CTCs obtained from patient blood offer clinically relevant insights into tumor metastasis and facilitate cancer diagnosis and treatment (Hong et al. 2016). Various studies to date have unequivocally highlighted the association between the abundance of CTCs in peripheral blood and poor disease prognosis (Cristofanilli et al. 2004; Danila et al. 2007; Giuliano et al. 2011; Rack et al. 2014; Bork et al. 2015; Tsai et al. 2016). Epithelial-to-mesenchymal transition (EMT) is believed to play a crucial role in metastasis. Under EMT, tumor epithelial cells acquire mesenchymal-like features for easy entry into the bloodstream (Bulfoni et al. 2016).

Recently developed platforms for CTC capture rely on diverse principles. These include antibody-based capture (Nagrath et al. 2007; Riethdorf et al. 2007; Stott et al. 2010), size exclusion (Xu et al. 2015), immune cell depletion (Ozkumur et al. 2013), and dielectrophoresis (Chiu et al. 2016). CellSearch, the only FDA-approved CTC capture platform, uses antibodies targeting the epithelial cell adhesion molecule (EPCAM) antigen for capturing CTC from patients' blood (Ignatiadis and Reinholz 2011; Habli et al. 2020; Iyer et al. 2020). The expression of epithelial markers like EPCAM and creatine kinase (CK) is used in affinity-based detection platforms to detect and count CTCs, but EMT arguably causes tumor cells to down-regulate or lose expression of canonical epithelial markers, thereby making them hard to recognize and capture while in the circulation (Iyer et al. 2020). As such, marker-based enrichment strategies are suboptimal for systematically charting heterogeneous CTC subpopulations (Miller et al. 2010; Farace et al. 2011; Wang et al. 2016). Various CTC capture platforms based on biophysical characteristics of cancer cells have been established in recent years (Ferreira et al. 2016; Gabriel et al. 2016; Cheng et al. 2019). Negative selection for the pan-leukocyte marker PTPRC has also been used as an alternative method. The promise of such antigen-agnostic platforms has not been explored adequately because the risk of immune cell contamination cannot be ruled out entirely (Ferreira et al. 2016; Gabriel et al. 2016). Isolation and molecular characterization of pure CTC populations by mRNA sequencing necessitate the development of precise analytic methods that are not based on epithelial markers and are able to spot leukocyte contamination.

The advent of single-cell RNA sequencing (scRNA-seq) has allowed in-depth, unsupervised analysis of CTC transcriptomes (Guo et al. 2015; Macosko et al. 2015; Kiselev et al. 2017, 2019a; Butler et al. 2018; Wolf et al. 2018; Chen et al. 2020; Ranjan et al. 2021). So far, most scRNA-seq studies involving CTCs have used marker-based approaches to zero in on CTC subpopulations. Marker-agnostic methods for CTC annotation are rare and often incapable of confirming the malignant identity of the cells. The major challenges involved are as follows: (1) high levels of intra- and intertumoral molecular diversity among malignant cells (Tirosh et al. 2016; Li et al. 2017); (2) the presence of CTCs in peripheral blood at an abysmally low concentration—one tumor cell within several millions of blood cells, even in patients with advanced metastatic disease (zero to 10 CTCs per mL of blood) (Alix-Panabières and Pantel 2013); (3) the fact that CTCs often undergo EMT, thereby disguising their epithelial markers (Mikolajczyk et al. 2011; Iyer et al. 2020); and (4) batch effect across scRNA-seq studies (Büttner et al. 2019; Kiselev et al. 2019a).

To overcome these challenges, we present unCTC, an R package for the unbiased characterization of CTC transcriptomes, in contrast with white blood cells (WBCs). unCTC features various standard and novel computational/statistical modules for clustering, copy number variation (CNV) inference, and marker-based characterization of CTC and non-CTC clusters obtained by analyzing the scRNA-seq data. For clustering, DDLK, a deep dictionary learning (DDL)–based method, is proposed. DDLK uses pathway scores at the single-cell level to accurately segregate CTC and WBC populations. With unCTC, we showed how in silico characterization of CTCs can strengthen marker-free CTC capturing and characterization. For this, we used the ClearCell Polaris workflow for size-based capture, immune cell depletion, and single-cell gene expression profiles of potential CTCs (Warkiani et al. 2014; Ramalingam et al. 2016). The unCTC workflow confers phenotypic identity on the captured cells through multifactorial analyses of the single-cell expression profiles.

Results

Overview of the unCTC workflow

Identification and characterization of CTC using scRNA-seq profiles are ever more challenging owing to the dynamic nature of the CTC phenotype. The unCTC workflow features a number of methods that help in the unbiased identification and characterization of single CTC transcriptomes. Clustering of scRNA-seq profiles is an important step toward this. Here we present a robust approach for clustering single-cell transcriptomes in a metaspace, spanning pathways, whose enrichment scores are computed on single-cell gene expression readouts. Single-cell expression data are typically sparse (Tian et al. 2019; Kiselev et al. 2019b). Pathway scores computed on gene sets alleviate this problem to a great extent (Li et al. 2017; Chawla et al. 2021), thereby assisting in the robust detection of cellular subtypes. For unsupervised clustering, each of the normalized and log-transformed expression vectors associated with CTCs is converted into a vector of pathway enrichment scores, calculated using gene set variation analysis (GSVA) (Hänzelmann et al. 2013). Such a transformation neutralizes batch effects (Kim et al. 2018) and unravels cellular heterogeneity from a rather functional/mechanistic point of view (Ding et al. 2019, 2020; Ramirez et al. 2020; Wang et al. 2020). DDLK incorporates the *k*-means clustering cost into the DDL framework. Shallow learning and data dependency are the main caveats of dictionary learning and deep learning, respectively. DDL aims to mitigate these challenges (Tariyal et al. 2016). DDLK is an example of semisupervised clustering that projects the single-cell gene expression data onto a range of well-understood biological pathways to obtain robust cellular clusters.

Although DDLK robustly identifies phenotypically similar cell groups from scRNA-seq data containing CTC expression profiles, cluster annotation may still remain elusive. To address this, we integrated the inferCNV function into the unCTC R package (RStudio Team 2022). inferCNV is an existing method, capable of inferring CNV from single-cell gene expression data (Couturier et al. 2020). Chromosomes in cancer cells undergo substantial aberration. inferCNV has been proven to be useful in capturing approximate CNV locations at single-cell resolution (Durante et al. 2020; Zhou et al. 2020). inferCNV works as a sounding board for cell-type characterization, especially zeroing in on the malignant origin of CTCs. Further, inferCNV along with cytoband information based on GRCh37 (Barrios and Prieto 2017) also pinpoints the precise position of chromosomal aberrations at the level of chromosomal arms, aiding in the identification of altered genes.

CTCs undergo EMT and other biophysical stress during their journey to distant organs. In this process, they partially lose their

epithelial phenotype. Univariate differential expression studies may turn out to be limitedly helpful in such scenarios. To circumvent this, unCTC allows the cumulative measurement of enrichment of a range of canonical markers, indicating malignant/epithelial/immune origin with the help of Stouffer's method (Stouffer et al. 1949). In our hands, such gene set–based approaches turn out to be fruitful in bolstering single-marker-based and inferred CNV-based characterization of cell groups. The complete unCTC workflow is outlined in Figure 1.

Marker-free capture of CTCs

Breast cancer is the most frequent type of cancer and one of the top causes of cancer-related mortality (Kamal et al. 2017). In 2020, breast cancer overtook lung cancer as the world's most common cancer. About 90% of breast malignancy–related fatalities are attributable to metastasis (Zhang et al. 2021). Breast cancer appears to be the form of cancer in which CTCs have been studied the most (Bidard et al. 2016). The expression of epithelial markers, including EPCAM and CK, has traditionally been used in affinity-dependent detection platforms to recognize and count CTCs, but these markers are down-regulated during EMT. Furthermore, most fluorescence-activated cell-sorting techniques face challenges owing to acute scarcity of CTCs, where the scarcity is typically fewer than one CTC per 10 mL of blood in nonmetastatic malignancies (Thery et al. 2019). Because of this inadequacy, a marker-free, robust method is necessary for detecting and enriching CTCs in a large pool of blood cells. Marker-free methods for isolating CTCs are appealing because they enable researchers to examine a greater number of CTCs that would otherwise be missed owing to variable or absent protein (label) marker expression on the CTC surfaces. It was possible to establish a marker-free method for isolating CTCs by integrating the ClearCell FX and Polaris systems (Iyer et al. 2020). As part of this, CTCs are enriched in two steps: size-based enrichment by ClearCell, followed by PTPRC (leukocyte marker) and CD31 (endothelial cell marker) based on negative depletion by Polaris (Fig. 2; Warkiani et al. 2014; Ramalingam et al. 2016). Using the ClearCell FX and Polaris systems, we collected 81 single CTCs from six women with breast cancer of three subtypes ($ER^-/PR^-/HER2^-$, $ER^+/PR^+/HER2^-$, and $ER^-/PR^-/HER2^+$) (Supplemental Table S1). Seventy-two CTCs finally qualified for the quality control criteria (Supplemental Table S2). In the subsequent sections, we illustrate unCTC-based characterization of these cells, in contrast to three other best practice integrative analysis methods, namely, Seurat (Hao et al. 2021), fastMNN (Haghverdi et al. 2018), and Harmony (Korsunsky et al. 2019). Execution details for Seurat, fastMNN, and Harmony can be found in Supplemental Note 1.

DDLK clustering leads to near-perfect segregation of CTCs and WBCs

The main aim of unCTC is to enable segregation of the CTC and WBC populations, obtained after unbiased microfluidic enrichment of CTCs in patient blood. This problem is fundamentally different from identifying CTC clusters or deciphering functional heterogeneity among single-cell transcriptomes, a method that typically requires unsupervised clustering of expression vectors. To cater to this objective, we chose to project expression vectors in a metaspace spanned by well-characterized biomolecular pathways using GSVA that, when supplied with selected pathways, convert given expression vectors into vectors comprising pathway enrichment scores (Hänzelmann et al. 2013). This is particularly advantageous and confers robustness in data integration tasks (Jin et al. 2014). Expression vectors, after conversion into vectors of pathway enrichment scores, are used for clustering the associated single-cell transcriptomes. Existing deep learning–based clustering techniques use stacked autoencoders (Peng et al. 2016; Xie et al. 2016; Yang et al. 2017; Fard et al. 2020) or their convolutional counterparts (Guo et al. 2017; Yang et al. 2019). Unlike DDL (Tariyal et al. 2016), the problem with autoencoders is that they have to estimate the parameters (encoder network + decoder network) twice. This leads to overfitting and general degradation of results. Prior studies have shown DDL to be the go-to framework for data-constrained scenarios (Fu et al. 2019; Mahdizadehaghdam et al. 2019; Tang et al. 2021) instead of conventional deep learning. DDLK (the clustering technique incorporated in unCTC) incorporates *k*-means clustering cost into the DDL framework, thereby enabling the clustering of data of all sizes. Presently most single-cell studies involve the integration of scRNA-seq data sets coming from different biological replicates, giving rise to significant batch effects (Sinha et al. 2019). Also, because of small amounts of starting RNA, single-cell data, even if it comes from a single chip, shows cell-to-cell technical variability. It is, therefore, imperative to ensure that a single-cell pipeline is robust to such variance factors. To validate this, we constructed a challenging multistudy (141 CTCs spanning breast, lung, and pancreatic cancers; 1037 WBCs) data set (Aceto et al. 2014; Ting et al. 2014; Yu et al. 2014; Sarioglu et al. 2015; Jordan et al. 2016; Velten et al. 2017; Zheng et al. 2017) for comparative assessment of unCTC, in contrast to Seurat, fastMNN, and Harmony (Supplemental Table S3). This integrative analysis task is referred to as Study 1. We subjected the data set to unCTC and three other best practice methods: Seurat, fastMNN, and Harmony. Two variants of Seurat were considered for the analysis: Vanilla Seurat and Integrative Seurat. Vanilla Seurat takes as input a single matrix, obtained by integrating individual scRNA-seq data sets based on common genes, whereas Integrative Seurat uses canonical correlation analysis (CCA)/reciprocal principal component analysis (RPCA)–based approaches for integration of multiple studies (Supplemental Note 1). Vanilla Seurat and unCTC managed to visually segregate CTCs and WBCs, whereas fastMNN and Harmony failed to separate the two categories (Fig. 3). Figure 3, A, B, D, E, G, H, J, and K, depicts the clustering performances by different methods in terms of CTC/WBC segregation. unCTC and Vanilla Seurat were able to find CTCs as part of different clusters (Fig. 3C, L). On the contrary, clusters returned by fastMNN and Harmony had a mixture of both cell types (Fig. 3F,I). PCA-, MNN-, and Harmony-based visualization for all methods can be found in Supplemental Figure S1.

Although clustering is the most popular means for unsupervised multivariate analysis of single cells, cell-lineage annotation requires the examination of marker-gene expression. Typically, univariate statistics are applied to user-selected canonical markers to confer lineage identity on CTCs and WBCs. Given the unpredictable expression dynamics of single markers in CTCs, it is often useful to measure combined up-regulation of tens of markers, reducing dependency on individual genes. To this end, we used Stouffer's method to combine expression levels of a range of markers associated with cell lineages of interest (Supplemental Table S4). Figure 4, A–C, highlights cluster-specific enrichment scores of genes related to immune genes, as well as breast elevated genes. Out of the three clusters retrieved by unCTC, cluster 2 showed the highest enrichment of epithelial markers. This finding is concordant with annotations sourced from the studies.

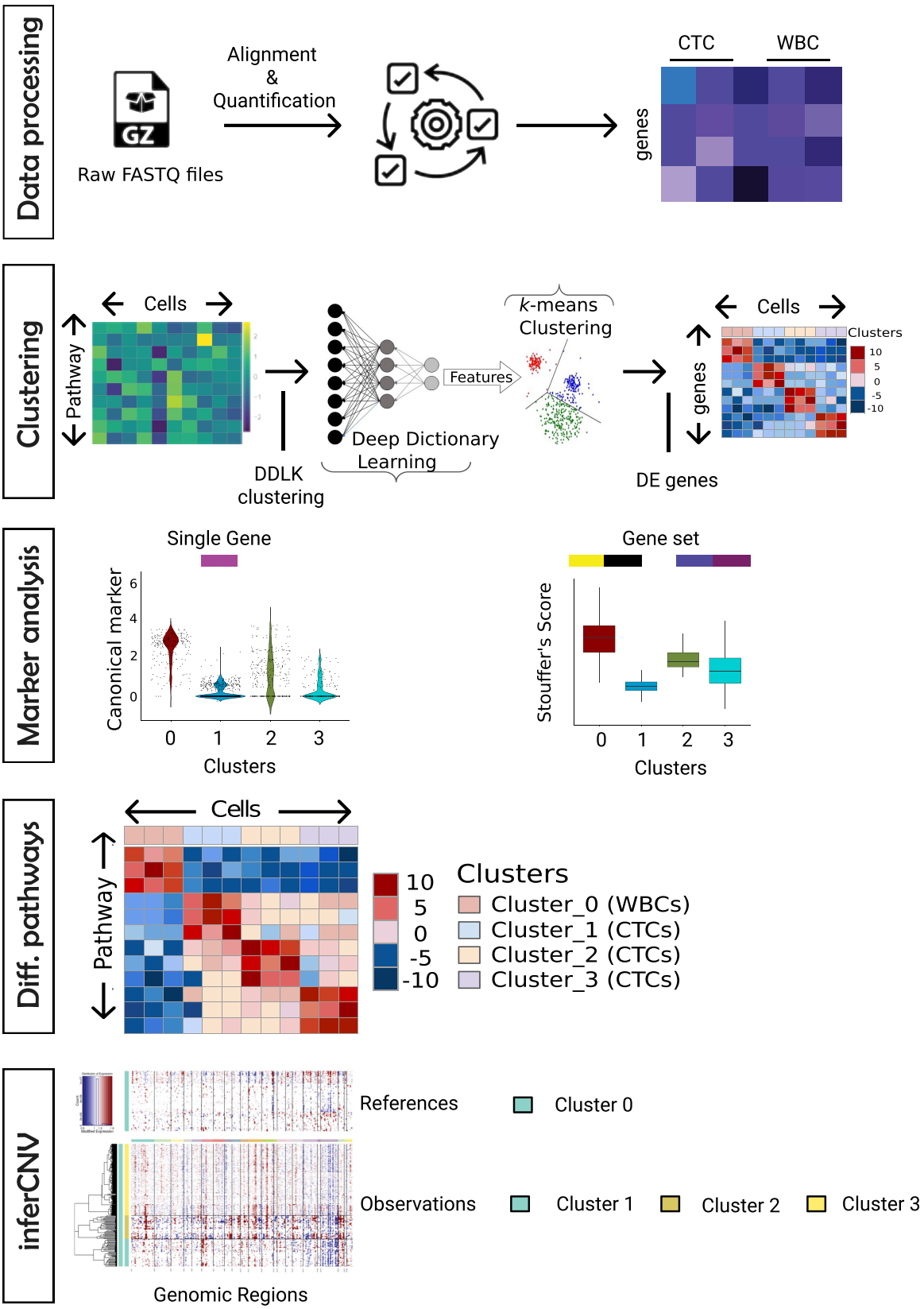


Figure 1. unCTC: a unified, end-to-end computational framework for marker-free characterization of CTCs. Schematic diagram depicting the analysis workflow, as well as the key methods supported by the unCTC R package. The first step involves processing raw FASTQ files to obtain the expression matrix. The novel DDLK clustering method is used to robustly cluster single CTC transcriptomes. Notably, DDLK works on pathway enrichment scores as opposed to expression values. Cluster-wise differential expression analysis is performed to gain insights into diverse CTC and WBC subtypes. Expression levels of well-known epithelial and immune markers are tracked to approximate broad cell-type identities. A similar analysis is also performed at the level of well-known gene sets/pathways. Furthermore, differential enrichment of pathway-specific genes can be analyzed to infer functional attributes. Finally, expression-based pseudo-CNV inference allows unbiased characterization of the identified clusters, thereby highlighting the malignant cells.

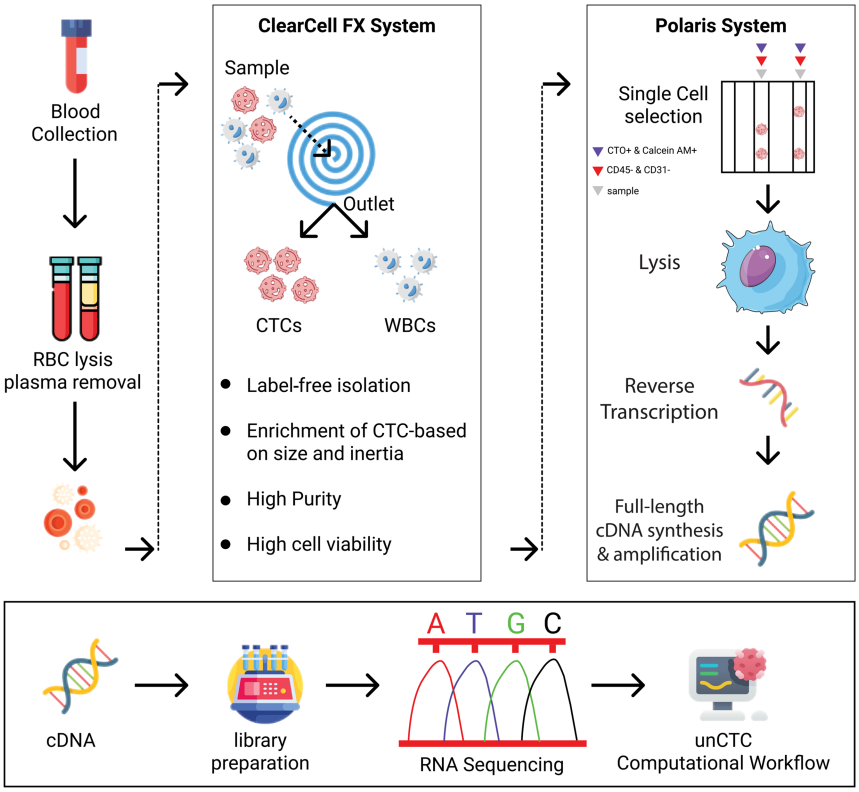


Figure 2. ClearCell FX and Polaris workflow for marker-free enrichment of CTCs. The schematic diagram depicts the key steps involved in the capture and isolation of CTCs using a two-pronged system. ClearCell FX uses a spiral chip to size-sort CTCs. Polaris performs single-cell capture and cDNA synthesis of potential CTCs after depletion of cells that are PTPRC/CD31-positive. Finally, cDNA thus received is subjected to library preparation and RNA sequencing.

unCTC recognizes CTCs selected by the ClearCell FX and Polaris workflow

We have recently demonstrated CTC characterization using supervised machine learning methods (Iyer et al. 2020). Given the dynamic nature of CTC phenotypes, it is however useful to characterize single CTC transcriptomes by unsupervised means. Further, classification-based characterization approaches are fallible in scenarios in which the obtained CTCs are of atypical phenotypes. unCTC alleviates this shortcoming by bringing to bear a spectrum of unbiased single-cell characterization tools. As an extended validation, we subjected the 72 filtered single-cell transcriptomes associated with potential CTCs captured by the ClearCell FX and Polaris workflow. These come from a total of six women entailing three major subtypes of breast cancer: $ER^-/PR^-/HER2^-$, $ER^+/PR^+/HER2^-$, and $ER^-/PR^-/HER2^+$. As a control, we also considered the CTC data set published by Ebright et al. (2020) that comprises 824 cells from 45 patients with breast cancer of the $ER^+/PR^+/HER2^-$ subtype. For WBCs, we considered 752 scRNA-seq profiles processed in two distinct runs using the Smart-seq2 protocol (Supplemental Table S3; Ding et al. 2020). This integrative analysis task is referred to as Study 2. Figure 5 summarizes the performance of all four methods, Vanilla Seurat, fastMNN, Harmony, and unCTC, in terms of their ability to segregate WBCs and CTCs. We found a mixture of CTCs and WBCs across most clusters with fastMNN and Harmony (Fig. 5D–I). Vanilla Seurat identified a number of clusters with CTCs alone; however, the embeddings appear to be confounded by batch effects (Fig. 5A–C). Clusters returned by unCTC grouped the CTCs into three clusters, whereas the WBCs were clumped into one large cluster (Fig. 5J–L). Notably, we found ClearCell FX and Polaris selected CTCs clustered with one of the ER^+ subgroups sourced from the study by Ebright et al. (2020). Supplemental Figure S2 depicts the PCA-, MNN-, and Harmony-based visualizations of the clusters, detected by all four methods. Out of the 72 CTCs that finally qualified the filtering criteria (obtained from ClearCell FX and Polaris workflow), ER^+ cells were most prevalent (54 out of 72). Among the rest, there were seven and 11 cells of the $HER2^+$ and triple-negative categories, respectively. One possible reason for not detecting $HER2^+$ and triple-negative CTCs as separate categories is their inadequate numbers, which makes it difficult for unCTC to retain relevant genes/pathways through several upstream filtering steps such as gene filtering and pathway selection.

Marker-dependent characterization of CTC clusters

Cell lineages are best understood through the enrichment of well-characterized lineage markers. Two approaches can be adopted for this: investigating differential expression for single markers and for marker panels. For the second study (comprising ClearCell FX and Polaris), we analyzed lineage identities for clusters identified by DDLK (Fig. 6A). Multiple well-known immune cell markers were spotted among the top 200 differentially up-regulated genes (Supplemental Table S5) among cells in cluster 0 that predominantly contains WBCs from the Ding et al. data set. These are *NKG7*, *PTPRC*, *PTPRCAP*, *IL32*, *CD74*, and *CD48*. The remaining clusters (clusters 1, 2, 3) comprise mostly CTCs (from the Poonia et al. and Ebright et al. data sets).

Cluster 1 among these is found to have elevated expression levels of integrins (*ITGA2B* and *ITGB5*). Integrins are principal adhesion molecules and play a central role in platelet function and hemostasis. Recent studies have postulated CTC–platelet interaction based on RNA extracts of single and clustered CTCs (Ting et al. 2014; Szczerba et al. 2019; Aceto 2020). CTCs constantly interact with factors in the blood such as platelets, circulating nucleic acids, and extracellular vesicles, which influence their molecular profiles (Ward et al. 2021). We observed elevated expression levels of the platelet degranulation markers *CLU* and *SPARC*, which are known for regulating *PF4* (Beck et al. 2019), a critical endocrine factor previously described to be associated with worse outcomes in patients with lung cancer (Pucci et al. 2016). *PF4* was also found to have elevated expression in cluster 1–specific cells. Cluster 1–specific CTCs showed elevated expression of numerous oncogenes with well-known roles in breast cancer progression. *CDKN1A* (Koch et al. 2020), *TIMP1* (Abreu et al. 2020), and *PGRMC1* (Clark et al. 2016) are notable among these. Cluster 2 that harbored

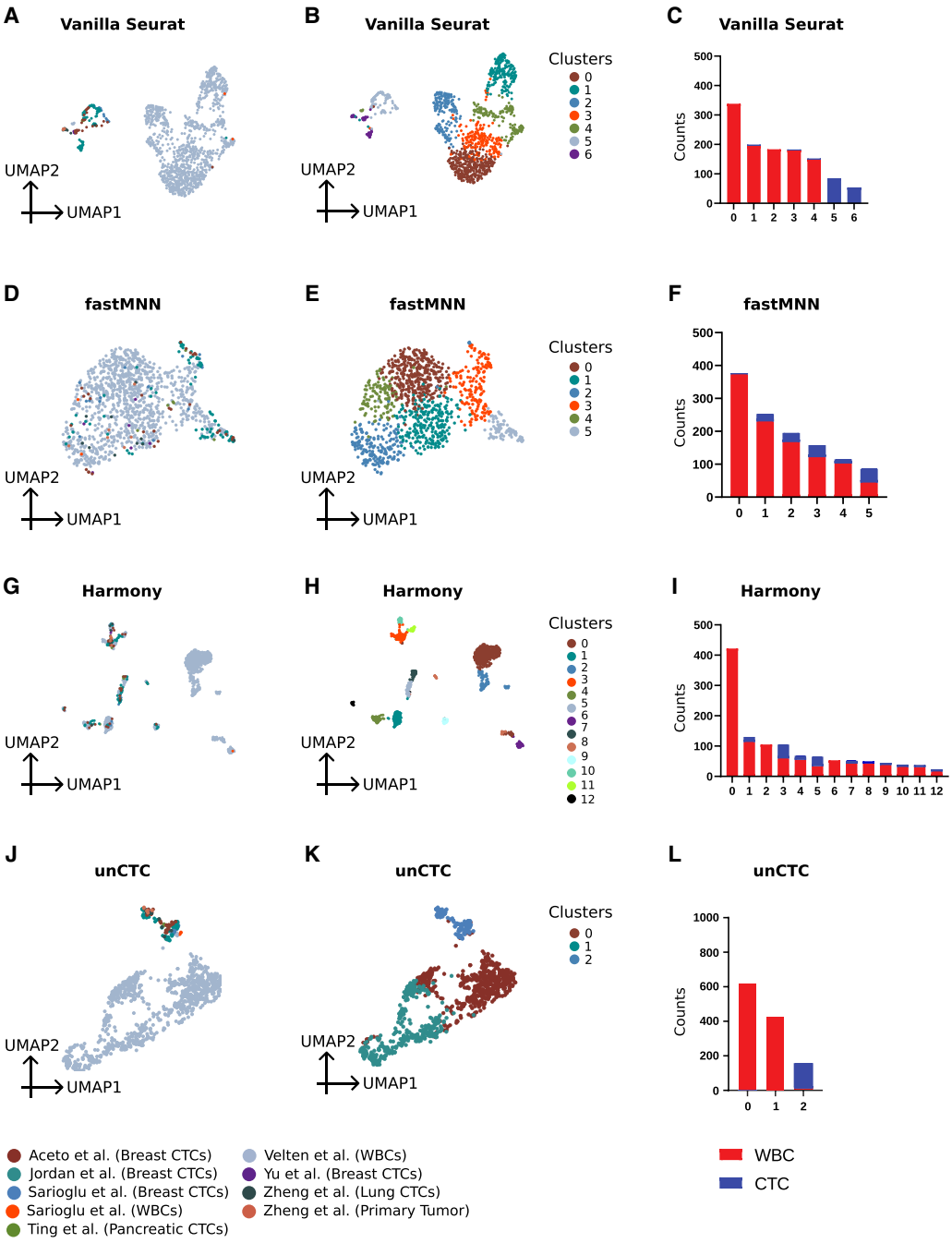


Figure 3. unCTC enables integrative analysis of CTCs and WBCs. A multistudy CTC/WBC data set was created to test the efficacy of unCTC alongside three best practice scRNA-seq analysis pipelines, namely, Vanilla Seurat, fastMNN, and Harmony. (*A–C*) Vanilla Seurat–based visualization, clustering, and cluster purity of CTCs and WBCs. (*D–F*) Similar figures for fastMNN. (*G–I*) Similar figures for Harmony. (*J–L*) Equivalent figures for unCTC.

the ClearCell/Polaris CTCs aside from CTCs from the Ebright et al. data set expressed a number of breast cancer–associated transcripts. Notable among these are—*IL10*, which drives breast cancer progression and proliferation (Sheikhpour et al. 2018); *BRIP1*, whose genotypic alteration increases breast cancer risk and elevated expression features invasive nature of the primary disease (Eelen et al. 2008); *IDO1*, which encodes a key immune checkpoint protein (Dill et al. 2018); and *POU5F1*, a cancer stem cell marker (Jin et al. 2019). Cluster 3 can be best characterized by the elevation of canonical epithelial markers such as *EPCAM, KRT18*, and *KRT19*. Tumor suppressor *SOD1* (Liu et al. 2020) was found to be up-regulated in these cells.

Functional analysis of the cluster-specific up-regulated genes was performed using IPA (Qiagen) (Krämer et al. 2014). Cluster 0 (WBC cluster)-specific up-regulated genes were largely unrelated to cancer, whereas the remaining clusters (CTC-specific) showed enrichment of cancer-associated pathways aligned with our above analysis of cluster marker genes (Fig. 6B–E). We also visualized

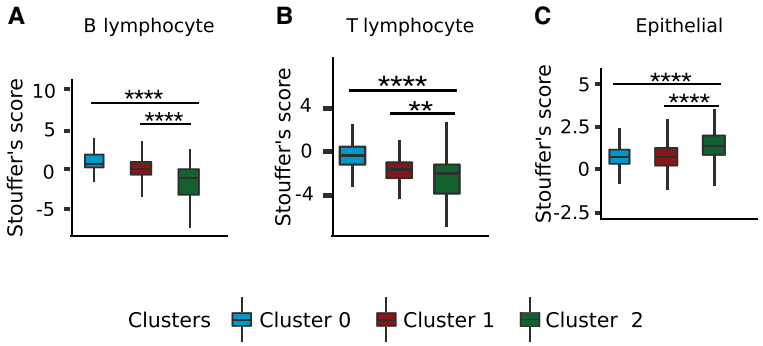


Figure 4. Cluster purity. (*A–C*) Boxplots depicting the distribution of Stouffer's scores computed based on known B cell, T cell, and epithelial cell markers, respectively, for cells in each of the unCTC identified clusters. Asterisks indicate that cluster differences are statistically significant.

selected relevant pathways that had differentially elevated enrichment in specific clusters (Supplemental Fig. S3). Cluster-specific differential enrichment of pathways was largely concordant with the insights gained based on analysis of differential expressed genes (Supplemental Table S6).

All ClearCell/Polaris CTCs clustered together with a fraction of Ebright CTCs. Notably, all Ebright CTCs are of the $ER^+/PR^+/HER2^-$ category. A large fraction of Poonia CTCs also belong to the same category (54 out of 72). Also, as discussed above, CTCs in cluster 2 were found to be enriched with breast cancer markers. This seems to be the dominant cause behind clustering of Poonia CTCs with a fraction of Ebright CTCs, leading to disguising the intersubtype heterogeneity in Poonia CTCs. To closely observe the heterogeneity among Poonia CTCs, we exclusively analyzed those using unCTC. We observed two subpopulations of $ER^+/PR^+/HER2^-$ CTCs and spatial localization of the triple-negative breast CTCs (Supplemental Fig. S4A). The Poonia et al. data set (ClearCell FX/Polaris system) encompasses all major breast cancer subtypes: $ER^-/PR^-/HER2^-$, $ER^-/PR^-/HER2^+$, and $ER^+/PR^+/HER2^-$. For each subtype, we identified the top 10 up-regulated genes (Supplemental Fig. S4B; Supplemental Table S7). The $ER^-/PR^-/HER2^+$ subtype showed elevated expression of *HEYL*. Notably, overexpression of the *HEYL* gene is known to be associated with poor survival in estrogen-negative breast cancer (Han et al. 2008). Another gene, *CREBL2*, has been reported to be expressed in a cell line with the $ER^+/PR^+/HER2^-$ subtype compared with another cell line with a basal subtype (Mellick et al. 2002). We found this gene to be up-regulated in the $ER^+/PR^+/HER2^-$ cells. *EIF3C*, which is known to promote proliferation (Zhao et al. 2017), showed elevated expression levels in CTCs originating from the triple-negative breast cancer (TNBC) subtype.

We performed cluster characterization based on single markers and marker panels. It is well known that owing to the loss of epithelial property, only a tiny percentage of CTCs would display conventional epithelial markers (Iyer et al. 2020). Because of EMT in CTCs and high dropout rates in scRNA-seq data, single markers may not show substantial differential expression; consequently, a combinatorial approach may be more beneficial than tracking the differential expression of individual genes. We curated markers from literature, and these markers are highly expressed in immune cells and breast epithelia (Supplemental Table S4). Stouffer's method (Stouffer et al. 1949) was used for combinatorial scoring of marker enrichment at a single-cell level. Scores thus obtained separated the WBC and CTC populations as expected (Supplemental Fig. S5A). We then tracked differential expressions of single markers. Cluster 0–specific cells were found to express high levels of leukocyte markers such as *PTPRC* and *NKG7*. *EPCAM* and *KRT18* showed relatively higher expression levels in cells specific to clusters 1, 2, and 3 (Supplemental Fig. S5B–E).

Expression-based CNV inference

Duplications and deletions that result in the addition or loss of significant chromosomal regions are referred to as CNVs. As proven by The Cancer Genome Atlas (The Cancer Genome Atlas Research Network et al. 2013) and the International Cancer Genome Consortium (Mafficini and Scarpa 2018), somatic CNVs, also known as copy number aberrations (CNAs), are prevalent in cancer. These CNAs are strongly linked to the onset, development, and metastasis of cancer (Sudmant et al. 2015; Jiang et al. 2016; Urrutia et al. 2018). With the growing popularity of single-cell sequencing of tumor microenvironments, expression-based CNV inference has become critical in zeroing in on malignant cells in a marker-independent manner (Tickle et al. 2019). The same strategy can be useful to characterize CTCs.

We subjected the Poonia et al. and Ebright et al. CTCs to inferCNV for CNV inference from single-cell expression data (Tickle et al. 2019). Separate analyses were performed for Poonia et al. CTCs and Ebright et al. CTCs because those come from different experimental pipelines and chemistries, and inferCNV is not equipped with batch effect correction. We used our internal PTPRC cell expression profiles from a healthy individual as a reference. More details pertaining to the experimental setting can be found in Supplemental Notes 2 and 3. In both studies, we identified several similar CNV patterns at the following chromosomal sites: Chr1p36, Chr1p36.11, Chr1q21.3, Chr3q29, Chr4p16.3, Chr5q35, Chr16q24, Chr17p13.3, Chr19q13.2, and Chr19q13.43 (Fig. 7; Supplemental Fig. S6). Chromosome 1q harbors both tumor suppressor genes and oncogenes and is linked with breast carcinogenesis (Bièche et al. 1995). Previous studies reported two altered regions in Chromosome 1q: the smallest commonly deleted and the overrepresented regions at 1q21-31 and 1q41-q44, respectively (Bièche et al. 1995; Lobo 2008; Privitera et al. 2021). Allelic loss in region 1q21–23 may be a valuable prognostic biological marker for the detection of local relapse in breast cancer (Gaki et al. 2000; Salahshourifar et al. 2015). The high frequency of loss of heterozygosity at 1q21.3, which plays a crucial role in tumor malignancy, could be one of the genetic markers used to diagnose malignancies (Yang et al. 2005; Salahshourifar et al. 2015). The 1p36 region of the chromosome is a known alteration hotspot (Ragnarsson et al. 1999). The *TP73* gene is situated at 1p36 and is homologous to *TP53*, which has been suggested as a potential tumor suppressor gene (Garnis et al. 2005). Amplification at 1p36.33 is associated with poor clinical outcomes (Ragnarsson et al. 1999; Bhosale et al. 2017). Previous studies have reported the presence of CTC-like genomic gains on Chromosome 19 at a low frequency in primary breast cancer (Kanwar et al. 2015). Furthermore, several studies have suggested that gains on Chromosome 19 may have a role in aggressive forms of breast cancer (Turner et al. 2010; Natrajan et al. 2012). The 19q13 region, in particular, is associated with copy number gains of signatures involved in dormancy and tumor aggressiveness in CTCs. Some of the genes involved in promoting EMT, invasion,

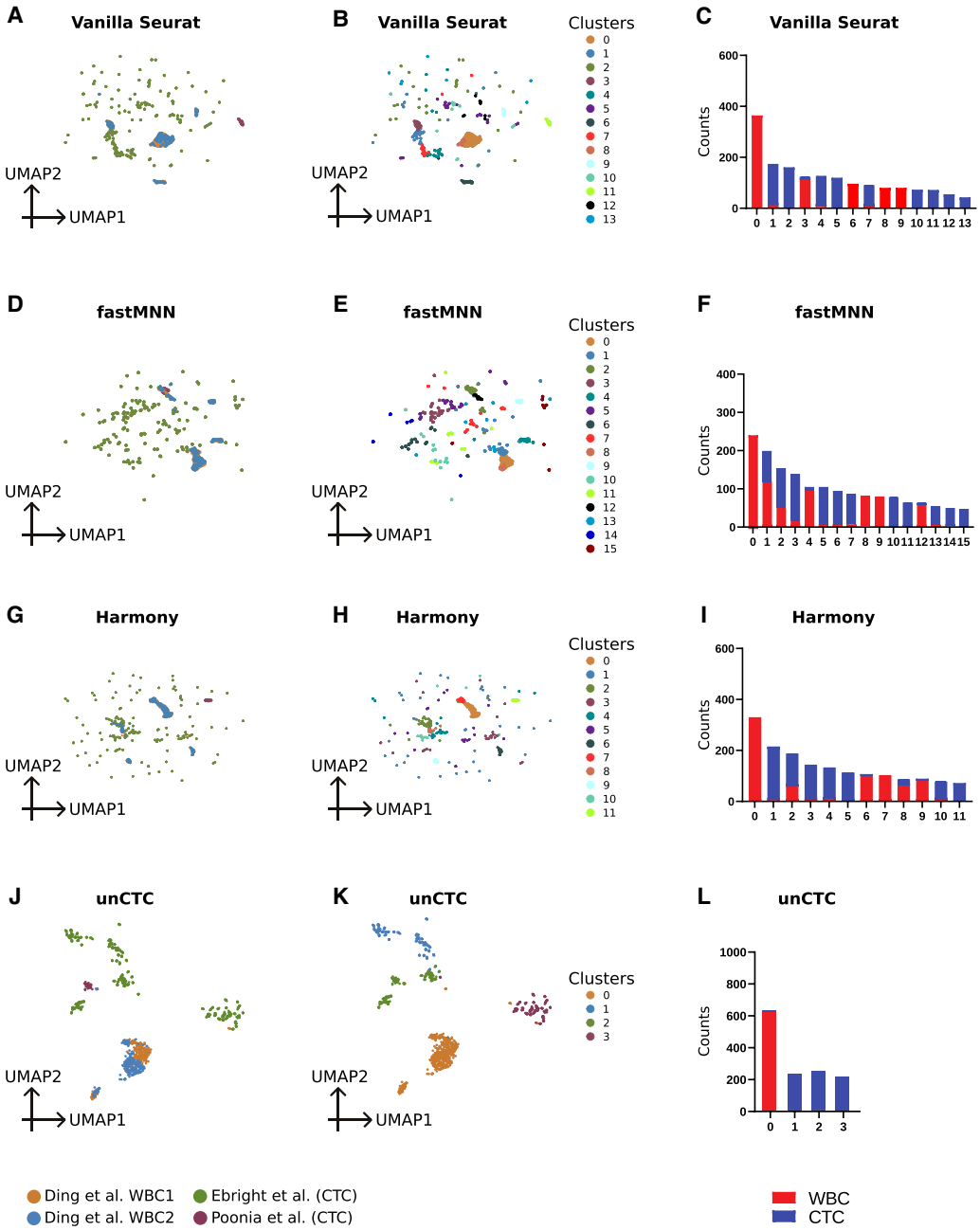


Figure 5. Clustering of CTCs obtained from ClearCell FX–Polaris system. (*A–I*) Visualization, clustering, and cluster purity of ClearCell FX–Polaris in presence of independent breast CTC and WBC scRNA-seq profiles, using Vanilla Seurat, fastMNN, and Harmony, respectively. (*J–L*) unCTC visualization, clustering, and cluster purity are depicted in equivalent figures. Notably, Vanilla Seurat, fastMNN, and Harmony failed to integrate CTCs from two different sources. unCTC accurately segregates CTCs and WBCs. CTCs obtained from ClearCell FX–Polaris system cocluster with breast CTCs from Ebright et al. data (Ebright et al. 2020).

and metastasis are *CEBPA* (19q13.11), *PAK4* (19q13.2), and *AKT2* (19q13.2) (Kanwar et al. 2015). The *APOBEC3B* gene, which is located on Chromosome 22q13.1, regulates breast cancer cell development by promoting ER transcriptional activity (Cescon et al. 2015). *APOBEC3B* copy number gain is associated with poor survival in patients with ER^+ breast cancer (Periyasamy et al. 2015; Murakami et al. 2021). A summary of each detected event along with the CNV state (1 = two copy losses, 2 = one copy loss, 3 = neutral, 4 = one copy gain, 5 = two copy gains, 6 = three or more copy gains) is given in Supplemental Tables S8 and S9, with source data identities.

Discussion

CTCs provide a window into their respective tumors of origin and cancer evolution. Most of the existing CTC enrichment methods are incomprehensive because they miss CTCs that do not express

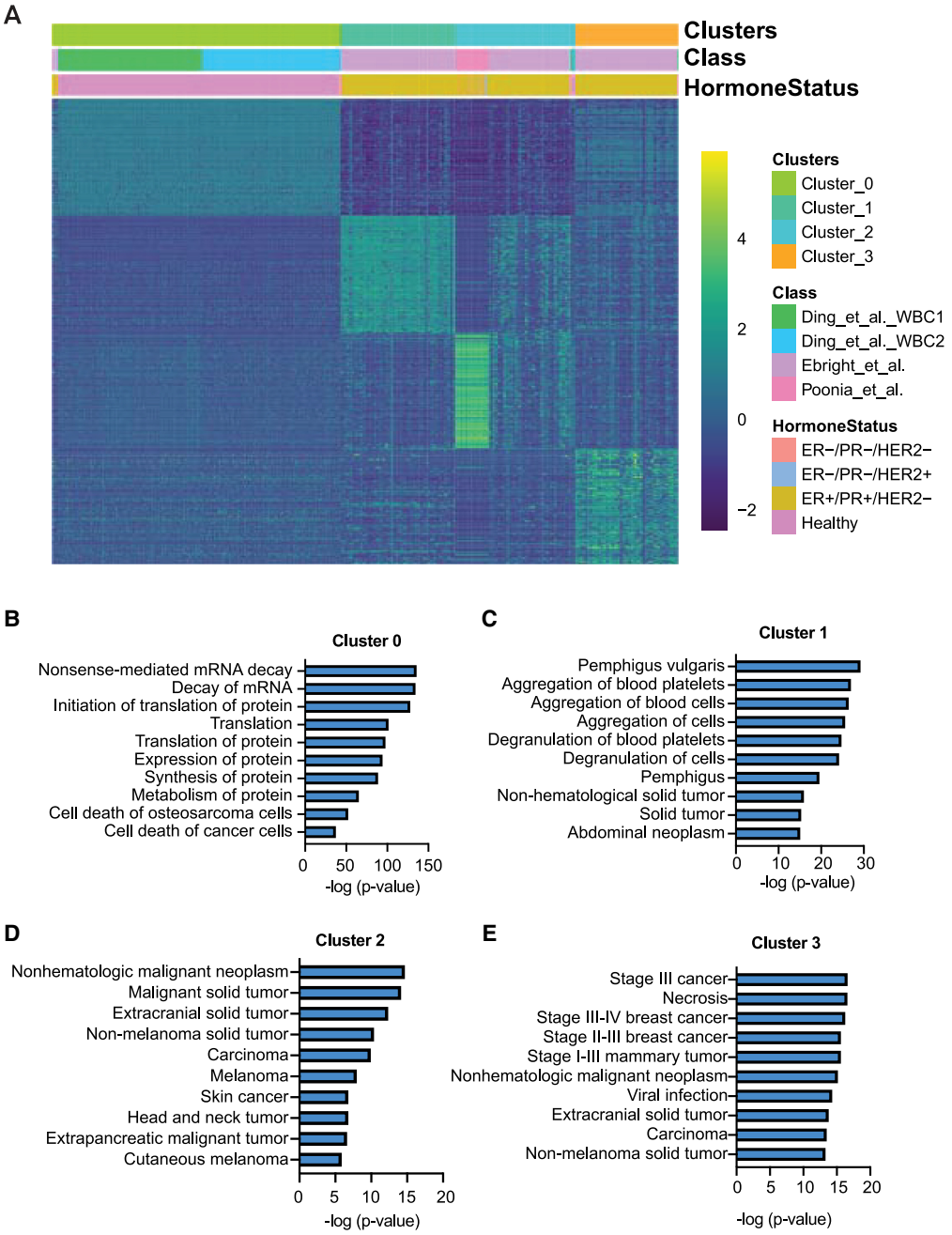


Figure 6. Functional annotations of the highly enriched CTC-associated genes identified using unCTC. (*A*) Heatmap depicting the expression of top 200 differentially elevated genes across four clusters detected by unCTC. Color bars indicate cluster identity, source data information, and molecular subcategories. (*B–E*) Bar plots depicting gene set enrichment using disease and functional annotation modules of IPA. The lists of cluster-specific differentially elevated genes can be found in Supplemental Table S5.

canonical epithelial markers. Single-cell expression studies allow the inspection of molecular profiles of CTC-rich cell populations obtained from an enrichment device. To date, there is no comprehensive computational resource that allows the identification of diverse/unknown CTC phenotypes from scRNA-seq data comprising CTCs and WBCs. unCTC enables this by providing a number of unsupervised and semisupervised means to interrogate CTC and WBC transcriptomes.

We demonstrated DDLK, a novel clustering approach that leverages pathway enrichment scores to yield robust grouping of single cells even when the data sets are sourced from disparate studies. This is particularly helpful because typical single-cell studies feature multiple replicates. It should be noted that DDLK is meant to discover broad groups in scRNA-seq data. It is neither tested nor expected to aid the discovery of heterogeneous subpopulations. For that, one can refer to our previous works describing the dropClust software suite (Sinha et al. 2018, 2019). With the help of unCTC, we could spot CTCs with unknown phenotypes. Expression-based CNV inference corroborated our findings with precise genomic locations,

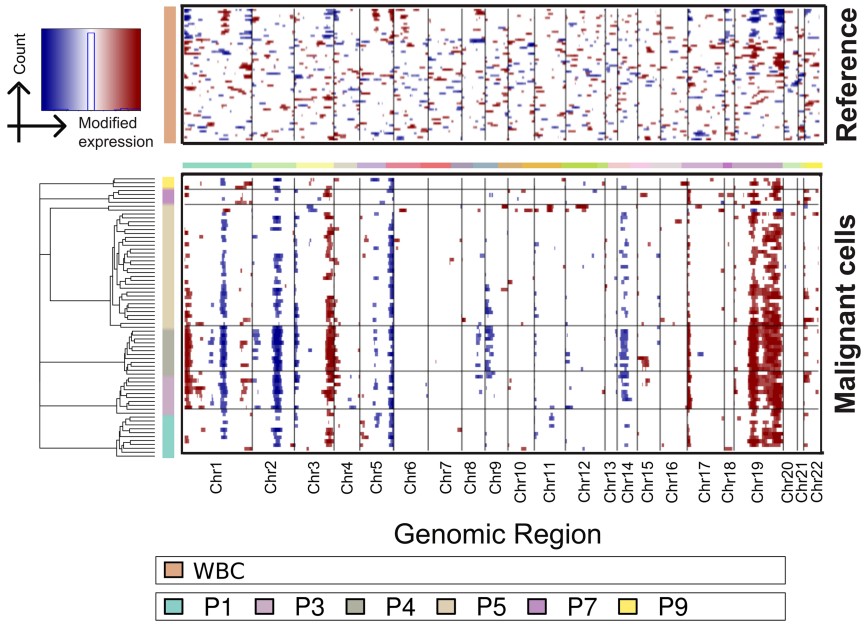


Figure 7. Expression-based inference of the CNV landscape across patient-wise malignant cells. The heatmap obtained from the inferCNV tool (Tickle et al. 2019) depicts the putative CNV landscape across six breast cancer patients while considering healthy WBCs as reference.

indicating amplifications/deletions that are previously known in breast cancers.

Tumor microenvironment profiling using scRNA-seq technologies is becoming increasingly popular. Similar to unbiased annotation of CTCs, a common challenge in tumor scRNA-seq data analysis is to distinguish between malignant and other nonmalignant stromal cell populations. Gene expression–based CNV inference is typically used as a rescue in these scenarios. However, in practice, one cannot guarantee the availability or expression-level detectability of the CNVs. To demonstrate the utility of unCTC in such scenarios, we analyzed scRNA-seq profiles from 18 primary head and neck squamous cell carcinoma (HNSCC) patients from a previously published study by Puram et al. (2017). unCTC clearly separated the malignant and nonmalignant populations, without any manual intervention (Supplemental Note 4; Supplemental Fig. S7).

In total, we compared unCTC with four state-of-the-art integrative single-cell analysis pipelines: Vanilla Seurat, Integrative Seurat (with CCA and RPCA variants), fastMNN, and Harmony (figures and supplemental notes associated with the benchmarking results can be found indexed in Supplemental Table S10). Vanilla Seurat managed to avoid mixing both cell types (Figs. 3, 5; Supplemental Note 1; Supplemental Figs. S1, S2). However, the low-dimensional embeddings appeared to be rather discrete for CTC subpopulations in Study 2 (Fig. 5; Supplemental Fig. S2). Of note, the Integrative Seurat variants (CCA and RPCA) struggled to stratify the CTC and WBC subpopulations (Supplemental Note 1; Supplemental Figs. S8, S9). For objective assessment of cluster qualities, we evaluated three frequently used metrics, namely, adjusted Rand index (ARI), normalized mutual information (NMI), and cluster purity for clusters produced by different methods under comparison. unCTC clearly outperformed the existing methods (Supplemental Figs. S1I, S2I, and S11I).

An integrative analysis method should ideally be agnostic of prenormalization strategy. This is a practical challenge faced by the bioinformatics community. We therefore tested the methods both for count (Study 1 and 2) and TPM-normalized data (Study 2). unCTC was found to be the only method to provide balanced performance in all cases, both in segregating WBC/CTC subpopulations and in providing coherent embeddings of cells (Figs. 3, 5; Supplemental Figs. S1, S2, S10, S11).

Upon exclusive analysis of Poonia et al. CTC data set, we could see spatial segregation of CTCs of the TNBC subcategory. TNBCs lack canonical surface markers, impeding their detectability in patient blood. As such, most CTC studies concentrate on the other breast cancer categories. The ClearCell FX–Polaris system used in this study could be potentially powerful in detecting CTCs in patients with triple-negative breast cancer, owing to its marker-agnostic approach. Of note, Ebright et al. CTCs were captured using the CTC-iChip technology (Ozkumur et al. 2013) while using epithelial/other cancer-specific markers to identify CTCs postenrichment. Such an approach could be unsuitable for detecting TNBCs owing to the lack of surface markers.

This study presents a unique experimental plus analysis workflow for marker-free CTC detection and characterization. On integrative analysis tasks, unCTC performed more consistently compared with Seurat, fastMNN, and Harmony. Notably, Vanilla Seurat was also capable of yielding reasonable separation of the cell types of interest. Although unCTC is robust to batch effects, it still needs improvement for adequately deciphering malignant cell heterogeneity. Our study is among the few that treat CTC transcriptome analysis as a unique challenge. Our study puts forth a unified computational framework and data sets, which together can serve as a baseline in this field of study.

CTCs are crucial biomarkers to monitor cancer progression and treatment response. Given the increasing throughput and sharply dropping price associated with single-cell expression profiling, we predict unCTC will play an important role in constructing cancer-specific molecular atlas of CTCs.

Methods

Description of data sets

We compiled two scRNA-seq data sets for comprehensive evaluation of unCTC. In the first study, we used seven distinct scRNA-seq data sets of CTCs and WBCs (Aceto et al. 2014; Ting et al. 2014; Yu et al. 2014; Sarioglu et al. 2015; Jordan et al. 2016; Velten et al. 2017; Zheng et al. 2017). Six of the seven data sets yielded 141 single CTCs. Two of the studies offered a total of 1037 WBCs. Notably, one of these data sets (obtained from the NCBI Gene Expression Omnibus [GEO; https://www.ncbi.nlm.nih.gov/geo/] under accession number GSE67939) contains both blood and CTC transcriptomes (Sarioglu et al. 2015). Another data set (GEO accession number GSE74639) contains 10 single CTCs and six single primary tumor cells (Zheng et al. 2017). The CTC data entail three cancer types: breast, lung, and

pancreatic (Supplemental Table S3). This data set was used to validate the unCTC potential for integrative analysis and clustering.

In the second study, we used three publicly accessible scRNA-seq data sets. The first data set consisted of 81 potential CTCs, enriched by the ClearCell FX and Polaris workflow. The CTCs were six patients with breast cancer of three subtypes: $ER^-/PR^-/HER2^-$, $ER^+/PR^+/HER2^-$, and $ER^-/PR^-/HER2^+$. These data are referred to as the Poonia et al. data set. In the second data set, as a control, we also considered 824 ER^+/PR^+ single CTCs isolated directly from whole-blood specimens of cancer patients using the CTC-iChip microfluidic system (the Ebright et al. data set) (Ebright et al. 2020). In the third data set, there were a total of 752 WBC expression profiles (processed in two different runs). This data set is referred to as the Ding et al. data set (Supplemental Table S3; Ding et al. 2020). We noted that, on average, the gene expression levels of the Poonia CTCs are found to be higher compared with the Ebright CTCs (Supplemental Fig. S12).

Sample collection

In total, 81 CTCs were collected from blood specimens of six breast cancer patients with distinct molecular subtypes. Out of these, 11 CTCs were obtained from one patient with TNBC, 57 CTCs from three patients with $ER^+/PR^+/HER2^-$ breast cancer, and 13 CTCs from three patients with $ER^-/PR^-/HER2^+$ subtype (Supplemental Table S1). All blood samples were collected from breast cancer patients at the National Cancer Center Singapore with the informed consent of all human participants and in accordance with institutional review board (IRB) guidelines (CIRB no. 2014/119/B). The SingHealth centralized institutional review board examined and approved the clinical sample collection protocols. The immunohistochemical (IHC) testing of estrogen receptor (ESR1, also known as ER), progesterone receptor (PGR, also known as PR), and erb-b2 receptor tyrosine kinase 2 (ERBB2, also known as HER2) status was conducted based on the latest guidelines of the American Society of Clinical Oncology and the College of American Pathologists.

CTC enrichment

To perform CTC enrichment, 9 mL of blood samples was collected in K3 EDTA blood collection tubes (Greiner Bio-One 455036). For each run, 6–8.5 mL of whole blood was processed. The red blood cells (RBCs) were first removed with the addition of the RBC lysis buffer (G-Biosciences) followed by incubation of 10 min at room temperature. After centrifugation, the lysed RBCs in the supernatant were removed. The nucleated cell pellet was suspended in a ClearCell resuspension buffer before CTC enrichment on the ClearCell FX system (Biolidics Limited) (Lee et al. 2018) per the manufacturer's instructions.

Immunofluorescence suspension staining

CTC-rich blood samples were centrifuged at 300*g* for 10 min and concentrated to 70 μL. The cell staining was performed with the addition of the following markers and antibodies for 1 h: CellTracker Orange (CTO; Thermo Fisher Scientific C34551), Calcein AM (Thermo Fisher Scientific L3224), PTPRC antibody conjugated with Alexa Fluor 647 (BioLegend 304020), and CD31 conjugated with Alexa Fluor 647 (BioLegend 303111). To improve the viability and RNA quality of the cells, 15 μL of RPMI with 10% FBS (Gibco) and 3 μL of RNase inhibitor (Thermo Fisher Scientific N8080119) were also added. 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. To achieve optimal buoyancy in an integrated fluidic circuit (IFC), 45 μL of CTCs was mixed with 30 μL cell suspension reagent (Fluidigm 101-0434) to achieve 75 μL of cell mix.

IFC operation

The Polaris IFC was first primed using the Fluidigm Polaris system to fill the control lines on the fluidic circuit, load cell capture beads, and block the inside of polydimethylsiloxane (PDMS) channels to avoid nonspecific absorption/adsorption of proteins. Then to capture and maintain the single cells in the sites, capture sites (48 sites) were preloaded with beads that are coupled on IFC to build a tightly packed bead column during the IFC priming process. After the priming stage, the cell mix (cells with suspension reagent) was laden in three inlets (25 μL each of cell mix) on the Polaris IFC and single CTO+ & Calcein AM+ & $PTPRC^-$ & $CD31^-$ cells were selected to capture sites. Finally, single-cell processing was achieved through template-switching mRNA-seq chemistry for full-length cDNA generation and preamplification on IFC. Supplemental Note 5 elaborates the steps involved in mRNA-seq library preparation and sequencing. The preprocessing steps for scRNA-seq data sets are outlined in Supplemental Note 6 (Supplemental Fig. S13).

Data integration, filtration, and normalization

RSEM software returns both read count data and TPM data (Poonia et al. and Ebright et al.) From these data sets, we discarded cells with a total read count of fewer than 50,000. As per this criterion, nine CTCs were removed from the Poonia et al. scRNA-seq data set (Supplemental Table S2). All 824 cells in the Ebright et al. data set qualified this criterion. The Poonia et al. and Ebright et al. data sets were integrated with the Ding et al. data set containing WBC expression profiles, and genes that are common across the data sets were retained. Further gene/cell filtering steps were implemented as follows. We eliminated cells with fewer than 1500 expressed genes (nonzero read count). On the other hand, we considered genes with nonzero expression in at least five cells. Linnorm normalization technique method was used with default parameters for single-cell normalization and batch correction (Yip et al. 2017). Normalized expression values are log-transformed after the addition of one as pseudo count. Linnorm normalization and log transformation are applied to both the count matrix and the TPM matrix.

Expression values to pathway enrichment scores

For computing gene set enrichment scores, we used the GSVA R software package (Hänzelmann et al. 2013). GSVA needs mainly two inputs: normalized and log-transformed expression matrix and gene sets. We used the C2 collection from MSigDB (Subramanian et al. 2005). This contains more than 6000 literature-curated gene sets. Before passing the gene set and expression matrix to the GSVA function, a filtering step is applied on gene sets that remove genes that are not present in the normalized expression matrix. We set *min.sz* as 10, *max.sz* as 500, *max.diff* as FALSE. Because the calculations for each gene set are independent of each other, we calculated enrichment scores in parallel. Here we used four parallel threads to speed up the computation (parallel.sz = 4).

DDLK clustering

Unsupervised clustering of GSVA enrichment scores was performed using *k*-means friendly DDL. To specify the optimal value of clusters for *k*-means clustering, the elbow method is used.

The popular way to express *k*-means clustering is via the following formulation:

$$\begin{aligned}&\sum_{i=1}^{k}\sum_{j=1}^{n} h_{ij}\|z_j-\mu_i\|_2^2\\ &h_{ij}\ =1,\quad \textit{if } x_j \in \textit{Cluster } i\\ &h_{ij}\ =0,\quad \textit{otherwise}\end{aligned} \tag{1}$$

where z_j denotes the jth sample and μ_i the ith cluster.

An alternate formulation for *k*-means is via matrix factorization (Bauckhage 2015):

$$\left\|Z-ZH^T(HH^T)^{-1}H\right\|_F^2, \tag{2}$$

where Z is the data matrix formed by stacking z_j's as columns, and H is the matrix of binary indicator variables h_{ij}. We prefer expressing *k*-means as Equation 2 in this work.

Because DDL is not a popular framework, we review it briefly. Dictionary learning (Tošić and Frossard 2011) learns a basis (D) such that the data (X) can be generated/synthesized from the coefficients (Z):

$$X = DZ. \tag{3}$$

The term dictionary learning is relatively new. The same problem has been known as matrix factorization in the past. One can see that Equation 3 is factoring the data matrix X into D and Z. In its most basic form, dictionary learning/matrix factorization is solved via the following:

$$\min_{D,Z}\|X-DZ\|_F^2, \tag{4}$$

where $\|\|_F^2$ is the squared Frobenius norm defined as the sum of the squares of all the terms in the matrix.

In DDL, instead of learning one layer of the dictionary, multiple layers are learned instead. This is expressed as

$$X = D_3\varphi(D_2\varphi(D_1Z)). \tag{5}$$

Here D_1, D_2, D_3 are three layers of dictionaries, and ϕ is the activation function between two layers. It is shown for three layers as an example; it can be more than three.

The solution to the unsupervised formulation is expressed as follows:

$$\min_{D_1,D_2,D_3,Z}\left\|X-D_3\varphi(D_2\varphi(D_1Z))\right\|_F^2. \tag{6}$$

In Equation 9, a greedy solution to Equation 6 was proposed. This was not optimal in the sense that there was feedback from shallow to deeper layers but not vice versa. To overcome this, the joint solution was proposed by Singhal and Majumdar (2018) based on the majorization minimization (MM) approach.

In this work, we will use the ReLU activation function for two reasons: (1) it is easier to incorporate as an optimization constraint, and (2) ReLU has been proven to have better function approximation capabilities. Therefore, our basic framework for DDL (with ReLU) will be expressed as follows:

$$\min_{D_1,D_2,D_3,Z,H}\ \|X-D_1D_2D_3Z\|_F^2\ \ s.t.\ D_2D_3\ Z\geq 0, D_3Z\geq 0, Z\geq 0. \tag{7}$$

ReLU activation

We propose to incorporate the *k*-means cost (Eq. 2) into the DDL formulation (Eq. 7). The basic idea is to use the features generated by DDL as inputs for clustering. However, instead of solving it piecemeal, we jointly optimize the following cost function:

$$\min_{D_1,D_2,D_3,Z,H}\underbrace{\|X-D_1D_2D_3Z\|_F^2}_{DDL}+\mu\underbrace{\left\|Z-ZH^T(HH^T)^{-1}H\right\|_F^2}_{k-means} \tag{8}$$

$$s.t.\quad D_2D_3Z\geq 0, D_3Z\geq 0, Z\geq 0\ .$$

We solve Equation 8 using alternating minimization. Initially, we ignore the nonnegativity constraints in Equation 8; later on, we will discuss how they can be handled. The updates for different variables are as follows:

$$\begin{aligned}&D_1\leftarrow\min_{D_1}\|X-D_1D_2D_3Z\|_F^2\\ &D_1^k=XZ_1^{\dagger},\quad \text{where } Z_1=D_2^{k-1}D_3Z^{k-1}.\end{aligned} \tag{9}$$

Here in the $Z_1^{\dagger}$, the superscript cross stands for Pseudoinverse.

$$\begin{aligned}&D_2\leftarrow\min_{D_2}\|X-D_1D_2D_3Z\|_F^2\\ &D_2^k=(D_1^k)^{\dagger}XZ_2,\quad \text{where } Z_2=D_3^{k-1}Z^{k-1}\end{aligned} \tag{10}$$

$$\begin{aligned}&D_3\leftarrow\min_{D_3}\|X-D_1D_2D_3Z\|_F^2\\ &D_3^k=(D_1^kD_2^k)^{\dagger}X(Z^{k-1})^{\dagger}\end{aligned} \tag{11}$$

$$Z\leftarrow\min_{Z}\|X-D_1D_2D_3Z\|_F^2+\mu\left\|Z-ZH^T(HH^T)^{-1}H\right\|_F^2. \tag{12}$$

To solve Z, we need to take the gradient of the expression in Equation 12 and equate it to zero. The derivation is given below.

$$\begin{aligned}&\nabla\left(\|X-D_1D_2D_3Z\|_F^2+\mu\left\|Z-ZH^T(HH^T)^{-1}H\right\|_F^2\right)=0\\ &\Rightarrow (D_1D_2D_3)^TX-(D_1D_2D_3)^T(D_1D_2D_3)^TZ-Z\left(\mu I-\mu H^T(HH^T)^{-1}H\right)=0\\ &\Rightarrow (D_1D_2D_3)^TX=(D_1D_2D_3)^T(D_1D_2D_3)^TZ+Z\left(\mu I-\mu H^T(HH^T)^{-1}H\right).\end{aligned}$$

The last step of the derivation implies that Z is a solution to Sylvester's equation of the form AX + XB = C. There are many efficient solvers for the same.

The final step is to update H. This is obtained by solving

$$H^k\leftarrow\min_{H}\left\|Z-ZH^T(HH^T)^{-1}H\right\|_F^2. \tag{13}$$

This is the *k*-means algorithm applied on Z.

In the derivation so far, we have not accounted for the ReLU nonnegativity constraints. Ideally, imposing the constraints would require solving them via forward–backward type splitting algorithms; such algorithms are iterative and hence would increase the run-time of the algorithm. We account for these constraints by simply putting the negative values in Z, Z_1, and Z_2 to zeroes in every iteration.

The algorithm is shown in a succinct fashion below. Once the convergence is reached, the clusters can be found from H. Because Equation 8 is a nonconvex function, we do not have any guarantees for convergence. We stop the iterations when the H does not change significantly in subsequent iterations.

Algorithm: DDL ± *k*-means
Initialize: $D_1^0, D_2^0, D_3^0, Z^0, H^0$
Repeat till convergence
  Update D_1^k, D_2^k, D_3^k *using* (9), (10), (11)
  Update Z^k *by solving Sylvester's eqn*
  Update H^k *by k – means clustering*
End

DEG identification

Differential genes between clusters obtained from DDLK clustering were calculated using the Limma (Ritchie et al. 2015) package with its Voom (Law et al. 2014) method. We first used the normalized expression matrix to construct a DEGlist obct. The DEGlist object was passed to the *calcNormFactors()* function of the edgeR R package (Robinson et al. 2010; RStudio Team 2022), while setting *normalisation factor = 1* and *method = none,* followed by Voom transformation. We reported the top up-regulated genes in each cluster, sorted by log fold change, and with a qualifying adjusted *P*-value cutoff of 0.05. Further functional analysis of the cluster-specific up-regulated genes was performed using ingenuity pathway analysis (IPA) (Krämer et al. 2014).

Differential pathways

The R package Limma was used to obtain differential pathways between clusters identified by the DDLK clustering (Ritchie et al. 2015). The moderated *t*-statistic was used for differential pathway analysis. Pathways with positive log fold change and adjusted *P*-value < 0.05 were considered specifically enriched in a cell group.

Combinatorial evaluation of lineage markers

Stouffer's method allows combining *Z*-scores across multiple variables to arrive at a single score indicating enrichment of a certain property (or phenotype) (Gupta et al. 2021). We used this to measure the enrichment of a spectrum of lineage indicating genes (breast epithelial and immune cell subtypes). This is particularly helpful for CTCs because single markers may not show adequate statistical significance for differential expression.

Software availability

Most of the updated code base and analysis pipeline can be accessed at GitHub (https://github.com/SaritaPoonia/unCTC). Relevant source codes are also provided as a Supplemental Code.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE186288 and GSE210651.

Competing interest statement

N.R. is an employee and stockholder of Fluidigm Corporation. A.A.B. and Y.F.L. are ex-employees of Biolidics and were stockholders in the company. D.S. is a stockholder in CareOnco Biotech and GentrepretR.

Acknowledgments

D.S. acknowledges the support of the iHub-Anubhuti-IIITD Foundation set up under the NM-ICPS scheme of the DST. D.S. also acknowledges the INSPIRE faculty grant awarded to D.S. from the Department of Science & Technology, India. S.P. acknowledges the University Grants Commission for supporting her PhD fellowship.

Author contributions: D.S. conceived the study with N.R. and A. Majumdar. S.P. performed all computational analyses with assistance from S.C., N.B., and P.R. A. Majumdar, along with D.S., supervised machine learning method development. S.P. implemented the method with help from A.G. N.R., J.W., and A.A.B. conceived the integration of ClearCell FX and Polaris. N.R. and Y.F.L. developed a marker-free workflow. Y.S.Y. provided the patient samples. Y.F.L. tested patient samples. G.A. assisted in interpreting results along with J.T. and A.M. G.A. ideated and improved the scientific illustrations. All authors contributed to writing and proofreading the manuscript.

References

Abreu M, Cabezas-Sainz P, Alonso-Alconada L, Ferreirós A, Mondelo-Macía P, Lago-Lestón RM, Abalo A, Díaz E, Palacios-Zambrano S, Rojo-Sebastian A, et al. 2020. Circulating tumor cells characterization revealed TIMP1 as a potential therapeutic target in ovarian cancer. *Cells* **9:** 1218. doi:10.3390/cells9051218

Aceto N. 2020. Bring along your friends: homotypic and heterotypic circulating tumor cell clustering to accelerate metastasis. *Biomed J* **43:** 18–23. doi:10.1016/j.bj.2019.11.002

Aceto N, Bardia A, Miyamoto DT, Donaldson MC, Wittner BS, Spencer JA, Yu M, Pely A, Engstrom A, Zhu H, et al. 2014. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **158:** 1110–1122. doi:10.1016/j.cell.2014.07.013

Alix-Panabières C, Pantel K. 2013. Circulating tumor cells: liquid biopsy of cancer. *Clin Chem* **59:** 110–118. doi:10.1373/clinchem.2012.194258

Barrios D, Prieto C. 2017. D3GB: an interactive genome browser for R, Python, and WordPress. *J Comput Biol* **24:** 447–449. doi:10.1089/cmb.2016.0213

Bauckhage C. 2015. k-means clustering is matrix factorization. arXiv:1512.07548 [stat.ML]. http://arxiv.org/abs/1512.07548

Beck TN, Boumber YA, Aggarwal C, Pei J, Thrash-Bingham C, Fittipaldi P, Vlasenkova R, Rao C, Borghaei H, Cristofanilli M, et al. 2019. Circulating tumor cell and cell-free RNA capture and expression analysis identify platelet-associated genes in metastatic lung cancer. *BMC Cancer* **19:** 603. doi:10.1186/s12885-019-5795-x

Bhosale PG, Cristea S, Ambatipudi S, Desai RS, Kumar R, Patil A, Kane S, Borges AM, Schäffer AA, Beerenwinkel N, et al. 2017. Chromosomal alterations and gene expression changes associated with the progression of leukoplakia to advanced gingivobuccal cancer. *Transl Oncol* **10:** 396–409. doi:10.1016/j.tranon.2017.03.008

Bidard F-C, Proudhon C, Pierga J-Y. 2016. Circulating tumor cells in breast cancer. *Mol Oncol* **10:** 418–430. doi:10.1016/j.molonc.2016.01.001

Bièche I, Champème MH, Lidereau R. 1995. Loss and gain of distinct regions of chromosome 1q in primary breast cancer. *Clin Cancer Res* **1:** 123–127.

Bittner KR, Jiménez JM, Peyton SR. 2020. Vascularized biomaterials to study cancer metastasis. *Adv Healthc Mater* **9:** e1901459. doi:10.1002/adhm.201901459

Bork U, Rahbari NN, Schölch S, Reissfelder C, Kahlert C, Büchler MW, Weitz J, Koch M. 2015. Circulating tumour cells and outcome in non-metastatic colorectal cancer: a prospective study. *Br J Cancer* **112:** 1306–1313. doi:10.1038/bjc.2015.88

Bulfoni M, Turetta M, Del Ben F, Di Loreto C, Beltrami AP, Cesselli D. 2016. Dissecting the heterogeneity of circulating tumor cells in metastatic breast cancer: going far beyond the needle in the haystack. *Int J Mol Sci* **17:** 1775. doi:10.3390/ijms17101775

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36:** 411–420. doi:10.1038/nbt.4096

Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. 2019. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* **16:** 43–49. doi:10.1038/s41592-018-0254-1

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Mills Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45:** 1113–1120. doi:10.1038/ng.2764

Cescon DW, Haibe-Kains B, Mak TW. 2015. *APOBEC3B* expression in breast cancer reflects cellular proliferation, while a deletion polymorphism is associated with immune activation. *Proc Natl Acad Sci* **112:** 2841–2846. doi:10.1073/pnas.1424869112

Chawla S, Samydurai S, Kong SL, Wu Z, Wang Z, Tam WL, Sengupta D, Kumar V. 2021. UniPath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. *Nucleic Acids Res* **49:** 1801. doi:10.1093/nar/gkab018

Chen L, Zhai Y, He Q, Wang W, Deng M. 2020. Integrating deep supervised, self-supervised and unsupervised learning for single-cell RNA-seq clustering and annotation. *Genes (Basel)* **11:** 792. doi:10.3390/genes11070792

Cheng Y-H, Chen Y-C, Lin E, Brien R, Jung S, Chen Y-T, Lee W, Hao Z, Sahoo S, Min Kang H, et al. 2019. Hydro-Seq enables contamination-free high-

throughput single-cell RNA-sequencing for circulating tumor cells. *Nat Commun* **10:** 2163. doi:10.1038/s41467-019-10122-2
Chiu T-K, Chou W-P, Huang S-B, Wang H-M, Lin Y-C, Hsieh C-H, Wu M-H. 2016. Application of optically-induced-dielectrophoresis in microfluidic system for purification of circulating tumour cells for gene expression analysis- Cancer cell line model. *Sci Rep* **6:** 32851. doi:10.1038/srep32851
Clark NC, Friel AM, Pru CA, Zhang L, Shioda T, Rueda BR, Peluso JJ, Pru JK. 2016. Progesterone receptor membrane component 1 promotes survival of human breast cancer cells and the growth of xenograft tumors. *Cancer Biol Ther* **17:** 262–271. doi:10.1080/15384047.2016.1139240
Couturier CP, Ayyadhury S, Le PU, Nadaf J, Monlong J, Riva G, Allache R, Baig S, Yan X, Bourgey M, et al. 2020. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat Commun* **11:** 3406. doi:10.1038/s41467-020-17186-5
Cristofanilli M, Budd GT, Ellis MJ, Stopeck A, Matera J, Miller MC, Reuben JM, Doyle GV, Allard WJ, Terstappen LWMM, et al. 2004. Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N Engl J Med* **351:** 781–791. doi:10.1056/NEJMoa040766
Danila DC, Heller G, Gignac GA, Gonzalez-Espinoza R, Anand A, Tanaka E, Lilja H, Schwartz L, Larson S, Fleisher M, et al. 2007. Circulating tumor cell number and prognosis in progressive castration-resistant prostate cancer. *Clin Cancer Res* **13:** 7053–7058. doi:10.1158/1078-0432.CCR-07-1506
Dill EA, Dillon PM, Bullock TN, Mills AM. 2018. IDO expression in breast cancer: an assessment of 281 primary and metastatic cases with comparison to PD-L1. *Mod Pathol* **31:** 1513–1522. doi:10.1038/s41379-018-0061-3
Ding H, Blair A, Yang Y, Stuart JM. 2019. Biological process activity transformation of single cell gene expression for cross-species alignment. *Nat Commun* **10:** 4899. doi:10.1038/s41467-019-12924-w
Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, Hughes TK, Wadsworth MH, Burks T, Nguyen LT, et al. 2020. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat Biotechnol* **38:** 737–746. doi:10.1038/s41587-020-0465-8
Durante MA, Rodriguez DA, Kurtenbach S, Kuznetsov JN, Sanchez MI, Decatur CL, Snyder H, Feun LG, Livingstone AS, Harbour JW. 2020. Single-cell analysis reveals new evolutionary complexity in uveal melanoma. *Nat Commun* **11:** 496. doi:10.1038/s41467-019-14256-1
Ebright RY, Lee S, Wittner BS, Niederhoffer KL, Nicholson BT, Bardia A, Truesdell S, Wiley DF, Wesley B, Li S, et al. 2020. Deregulation of ribosomal protein expression and translation promotes breast cancer metastasis. *Science* **367:** 1468–1473. doi:10.1126/science.aay0939
Eelen G, Vanden Bempt I, Verlinden L, Drijkoningen M, Smeets A, Neven P, Christiaens MR, Marchal K, Bouillon R, Verstuyf A. 2008. Expression of the BRCA1-interacting protein *Brip1/BACH1/FANCJ* is driven by E2F and correlates with human breast cancer malignancy. *Oncogene* **27:** 4233–4241. doi:10.1038/onc.2008.51
Farace F, Massard C, Vimond N, Drusch F, Jacques N, Billiot F, Laplanche A, Chauchereau A, Lacroix L, Planchard D, et al. 2011. A direct comparison of CellSearch and ISET for circulating tumour-cell detection in patients with metastatic carcinomas. *Br J Cancer* **105:** 847–853. doi:10.1038/bjc.2011.294
Fard MM, Thonet T, Gaussier E. 2020. Deep *k*-means: jointly clustering with *k*-means and learning representations. *Pattern Recognit Lett* **138:** 185–192. doi:10.1016/j.patrec.2020.07.028
Ferreira MM, Ramani VC, Jeffrey SS. 2016. Circulating tumor cell technologies. *Mol Oncol* **10:** 374–394. doi:10.1016/j.molonc.2016.01.007
Follain G, Osmani N, Azevedo AS, Allio G, Mercier L, Karreman MA, Solecki G, Garcia Leòn MJ, Lefebvre O, Fekonja N, et al. 2018. Hemodynamic forces tune the arrest, adhesion, and extravasation of circulating tumor cells. *Dev Cell* **45:** 33–52.e12. doi:10.1016/j.devcel.2018.02.015
Fu T, Hoang TN, Xiao C, Sun J. 2019. DDL: deep dictionary learning for predictive phenotyping. *IJCAI* **2019:** 5857–5863. doi:10.24963/ijcai.2019/812.
Gabriel MT, Calleja LR, Chalopin A, Ory B, Heymann D. 2016. Circulating tumor cells: a review of non–EpCAM-based approaches for cell enrichment and isolation. *Clin Chem* **62:** 571–581. doi:10.1373/clinchem.2015.249706
Gajewski TF, Schreiber H, Fu Y-X. 2013. Innate and adaptive immune cells in the tumor microenvironment. *Nat Immunol* **14:** 1014–1022. doi:10.1038/ni.2703
Gaki V, Tsopanomichalou M, Sourvinos G, Tsiftsis D, Spandidos DA. 2000. Allelic loss in chromosomal region 1q21–23 in breast cancer is associated with peritumoral angiolymphatic invasion and extensive intraductal component. *Eur J Surg Oncol* **26:** 455–460. doi:10.1053/ejso.1999.0921
Garnis C, Campbell J, Davies JJ, Macaulay C, Lam S, Lam WL. 2005. Involvement of multiple developmental genes on chromosome 1p in lung tumorigenesis. *Hum Mol Genet* **14:** 475–482. doi:10.1093/hmg/ddi043
Giuliano M, Giordano A, Jackson S, Hess KR, De Giorgi U, Mego M, Handy BC, Ueno NT, Alvarez RH, De Laurentiis M, et al. 2011. Circulating tumor cells as prognostic and predictive markers in metastatic breast cancer patients receiving first-line systemic treatment. *Breast Cancer Res* **13:** R67. doi:10.1186/bcr2907
Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. 2015. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. *PLoS Comput Biol* **11:** e1004575. doi:10.1371/journal.pcbi.1004575
Guo X, Liu X, Zhu E, Yin J. 2017. Deep clustering with convolutional autoencoders. In *International conference on neural information processing*, pp. 373–382. Springer, Cham. doi:10.1007/978-3-319-70096-0_39
Gupta K, Mohanty SK, Mittal A, Kalra S, Kumar S, Mishra T, Ahuja J, Sengupta D, Ahuja G. 2021. The cellular basis of loss of smell in 2019-nCoV-infected individuals. *Brief Bioinform* **22:** 873–881. doi:10.1093/bib/bbaa168
Habli Z, AlChamaa W, Saab R, Kadara H, Khraiche ML. 2020. Circulating tumor cell detection technologies and clinical utility: challenges and opportunities. *Cancers (Basel)* **12:** 1930. doi:10.3390/cancers12071930
Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36:** 421–427. doi:10.1038/nbt.4091
Han L, Lee JS, Mori T, Zhang H, Landberg G, Kallioniemi A, Argani P, Sukumar S. 2008. HEYL, an overexpressed gene in breast cancer, functions as a novel negative regulator of TGF-β pathway. *Cancer Res* **68:** 5201.
Hänzelmann S, Castelo R, Guinney J. 2013. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14:** 7. doi:10.1186/1471-2105-14-7
Hao Y, Hao S, Andersen-Nissen E, Mauck WM III, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184:** 3573–3587.e29. doi:10.1016/j.cell.2021.04.048
Hong Y, Fang F, Zhang Q. 2016. Circulating tumor cell clusters: what we know and what we expect (review). *Int J Oncol* **49:** 2206–2216. doi:10.3892/ijo.2016.3747
Ignatiadis M, Reinholz M. 2011. Minimal residual disease and circulating tumor cells in breast cancer. *Breast Cancer Res* **13:** 222. doi:10.1186/bcr2906
Iyer A, Gupta K, Sharma S, Hari K, Lee YF, Ramalingam N, Yap YS, West J, Bhagat AA, Subramani BV, et al. 2020. Integrative analysis and machine learning based characterization of single circulating tumor cells. *J Clin Med Res* **9:** 1206. doi:10.3390/jcm9041206
Jiang Y, Qiu Y, Minn AJ, Zhang NR. 2016. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc Natl Acad Sci* **113:** E5528–E5537. doi:10.1073/pnas.1522203113
Jin L, Zuo X-Y, Su W-Y, Zhao X-L, Yuan M-Q, Han L-Z, Zhao X, Chen Y-D, Rao S-Q. 2014. Pathway-based analysis tools for complex diseases: a review. *Genomics Proteomics Bioinformatics* **12:** 210–220. doi:10.1016/j.gpb.2014.10.002
Jin X, Li Y, Guo Y, Jia Y, Qu H, Lu Y, Song P, Zhang X, Shao Y, Qi D, et al. 2019. ERα is required for suppressing OCT4-induced proliferation of breast cancer cells via DNMT1/ISL1/ERK axis. *Cell Prolif* **52:** e12612. doi:10.1111/cpr.12612
Jordan NV, Bardia A, Wittner BS, Benes C, Ligorio M, Zheng Y, Yu M, Sundaresan TK, Licausi JA, Desai R, et al. 2016. HER2 expression identifies dynamic functional states within circulating breast cancer cells. *Nature* **537:** 102–106. doi:10.1038/nature19328
Kamal M, Razaq W, Leslie M, Adhikari S, Tanaka T. 2017. Circulating tumor cells in breast cancer: a potential liquid biopsy. In *Breast cancer* (ed. Van Pham P). IntechOpen, Rijeka, Croatia.
Kanwar N, Hu P, Bedard P, Clemons M, McCready D, Done SJ. 2015. Identification of genomic signatures in circulating tumor cells from breast cancer. *Int J Cancer* **137:** 332–344. doi:10.1002/ijc.29399
Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, Crosetto N, Foukakis T, Navin NE. 2018. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* **173:** 879–893.e13. doi:10.1016/j.cell.2018.03.041
Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14:** 483–486. doi:10.1038/nmeth.4236
Kiselev VY, Andrews TS, Hemberg M. 2019a. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* **20:** 273–282. doi:10.1038/s41576-018-0088-9
Kiselev VY, Andrews TS, Hemberg M. 2019b. Publisher Correction: Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* **20:** 310. doi:10.1038/s41576-019-0095-5
Kitamura T. 2018. A negative regulator of metastasis promoting macrophages. *J Emerg Crit Care Med* **2:** 56. doi:10.21037/jeccm.2018.06.01

Koch C, Kuske A, Joosse SA, Yigit G, Sflomos G, Thaler S, Smit DJ, Werner S, Borgmann K, Gärtner S, et al. 2020. Characterization of circulating breast cancer cells with tumorigenic and metastatic capacity. *EMBO Mol Med* **12:** e11908. doi:10.15252/emmm.201911908

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P-R, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* **16:** 1289–1296. doi:10.1038/s41592-019-0619-0

Krämer A, Green J, Pollard J Jr, Tugendreich S. 2014. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30:** 523–530. doi:10.1093/bioinformatics/btt703

Krebs MG, Metcalf RL, Carter L, Brady G, Blackhall FH, Dive C. 2014. Molecular analysis of circulating tumour cells: biology and biomarkers. *Nat Rev Clin Oncol* **11:** 129–144. doi:10.1038/nrclinonc.2013.253

Kwa M, Esteva FJ. 2018. Detection and clinical implications of occult systemic micrometastatic breast cancer. *The Breast* 858–866.e3. doi:10.1016/b978-0-323-35955-9.00066-0

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15:** R29. doi:10.1186/gb-2014-15-2-r29

Lee Y, Guan G, Bhagat AA. 2018. ClearCell® FX, a label-free microfluidics technology for enrichment of viable circulating tumor cells. *Cytometry A* **93:** 1251–1254. doi:10.1002/cyto.a.23507

Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, et al. 2017. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* **49:** 708–718. doi:10.1038/ng.3818

Liu Y, Cao X. 2016. Immunosuppressive cells in tumor immune escape and metastasis. *J Mol Med* **94:** 509–522. doi:10.1007/s00109-015-1376-x

Liu S, Li B, Xu J, Hu S, Zhan N, Wang H, Gao C, Li J, Xu X. 2020. SOD1 promotes cell proliferation and metastasis in non-small cell lung cancer via an miR-409-3p/SOD1/SETDB1 epigenetic regulatory feedforward loop. *Front Cell Dev Biol* **8:** 213. doi:10.3389/fcell.2020.00213

Lobo I. 2008. Chromosome abnormalities and cancer genetics. *Nat Education* **1:** 68.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161:** 1202–1214. doi:10.1016/j.cell.2015.05.002

Mafficini A, Scarpa A. 2018. Genomic landscape of pancreatic neuroendocrine tumours: the International Cancer Genome Consortium. *J Endocrinol* **236:** R161–R167. doi:10.1530/JOE-17-0560

Mahdizadehaghdam S, Panahi A, Krim H, Dai L. 2019. Deep dictionary learning: a PARametric NETwork approach. *IEEE Trans Image Process* **28:** 4790–4802. doi:10.1109/TIP.2019.2914376

Mathers CD. 2020. History of global burden of disease assessment at the World Health Organization. *Arch Public Health* **78:** 77. doi:10.1186/s13690-020-00458-3

McAllister SS, Weinberg RA. 2014. The tumour-induced systemic environment as a critical regulator of cancer progression and metastasis. *Nat Cell Biol* **16:** 717–727. doi:10.1038/ncb3015

Mellick AS, Day CJ, Weinstein SR, Griffiths LR, Morrison NA. 2002. Differential gene expression in breast cancer cell lines and stroma–tumor differences in microdissected breast cancer biopsies revealed by display array analysis. *Int J Cancer* **100:** 172–180. doi:10.1002/ijc.10451

Mikolajczyk SD, Millar LS, Tsinberg P, Coutts SM, Zomorrodi M, Pham T, Bischoff FZ, Pircher TJ. 2011. Detection of EpCAM-negative and cytokeratin-negative circulating tumor cells in peripheral blood. *J Oncol* **2011:** 252361. doi:10.1155/2011/252361

Miller MC, Doyle GV, Terstappen LWMM. 2010. Significance of circulating tumor cells detected by the CellSearch system in patients with metastatic breast colorectal and prostate cancer. *J Oncol* **2010:** 617421. doi:10.1155/2010/617421

Murakami F, Tsuboi Y, Takahashi Y, Horimoto Y, Mogushi K, Ito T, Emi M, Matsubara D, Shibata T, Saito M, et al. 2021. Short somatic alterations at the site of copy number variation in breast cancer. *Cancer Sci* **112:** 444–453. doi:10.1111/cas.14630

Nagrath S, Sequist LV, Maheswaran S, Bell DW, Irimia D, Ulkus L, Smith MR, Kwak EL, Digumarthy S, Muzikansky A, et al. 2007. Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* **450:** 1235–1239. doi:10.1038/nature06385

Natrajan R, Mackay A, Wilkerson PM, Lambros MB, Wetterskog D, Arnedos M, Shiu K-K, Geyer FC, Langerød A, Kreike B, et al. 2012. Functional characterization of the 19q12 amplicon in grade III breast cancers. *Breast Cancer Res* **14:** R53. doi:10.1186/bcr3154

Ozkumur E, Shah AM, Ciciliano JC, Emmink BL, Miyamoto DT, Brachtel E, Yu M, Chen P-I, Morgan B, Trautwein J, et al. 2013. Inertial focusing for tumor antigen-dependent and -independent sorting of rare circulating tumor cells. *Sci Transl Med* **5:** 179ra47. doi:10.1126/scitranslmed.3005616

Peng X, Xiao S, Feng J, Yau W-Y, Yi Z. 2016. Deep subspace clustering with sparsity prior. In *IJCAI'16: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1925–1931. AAAI Press, Palo Alto, CA. doi:10.5555/3060832.3060890

Periyasamy M, Patel H, Lai C-F, Nguyen VTM, Nevedomskaya E, Harrod A, Russell R, Remenyi J, Ochocka AM, Thomas RS, et al. 2015. APOBEC3B-mediated cytidine deamination is required for estrogen receptor action in breast cancer. *Cell Rep* **13:** 108–121. doi:10.1016/j.celrep.2015.08.066

Privitera AP, Barresi V, Condorelli DF. 2021. Aberrations of chromosomes 1 and 16 in breast cancer: a framework for cooperation of transcriptionally dysregulated genes. *Cancers (Basel)* **13:** 1585. doi:10.3390/cancers13071585

Pucci F, Rickelt S, Newton AP, Garris C, Nunes E, Evavold C, Pfirschke C, Engblom C, Mino-Kenudson M, Hynes RO, et al. 2016. PF4 promotes platelet production and lung cancer growth. *Cell Rep* **17:** 1764–1772. doi:10.1016/j.celrep.2016.10.031

Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, et al. 2017. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171:** 1611–1624.e24. doi:10.1016/j.cell.2017.10.044

Rack B, Schindlbeck C, Jückstock J, Andergassen U, Hepp P, Zwingers T, Friedl TWP, Lorenz R, Tesch H, Fasching PA, et al. 2014. Circulating tumor cells predict survival in early average-to-high risk breast cancer patients. *J Natl Cancer Inst* **106:** dju066. doi:10.1093/jnci/dju066

Ragnarsson G, Eiriksdottir G, Johannsdottir JT, Jonasson JG, Egilsson V, Ingvarsson S. 1999. Loss of heterozygosity at chromosome 1p in different solid human tumours: association with survival. *Br J Cancer* **79:** 1468–1474. doi:10.1038/sj.bjc.6690234

Ramalingam N, Fowler B, Szpankowski L, Leyrat AA, Hukari K, Maung MT, Yorza W, Norris M, Cesar C, Shuga J, et al. 2016. Fluidic logic used in a systems approach to enable integrated single-cell functional analysis. *Front Bioeng Biotechnol* **4:** 70. doi:10.3389/fbioe.2016.00070

Ramirez AK, Dankel SN, Rastegarpanah B, Cai W, Xue R, Crovella M, Tseng Y-H, Kahn CR, Kasif S. 2020. Single-cell transcriptional networks in differentiating preadipocytes suggest drivers associated with tissue heterogeneity. *Nat Commun* **11:** 2117. doi:10.1038/s41467-020-16019-9

Ranjan B, Schmidt F, Sun W, Park J, Honardoost MA, Tan J, Arul Rayan N, Prabhakar S. 2021. scConsensus: combining supervised and unsupervised clustering for cell type identification in single-cell RNA sequencing data. *BMC Bioinformatics* **22:** 186. doi:10.1186/s12859-021-04028-4

Riethdorf S, Fritsche H, Müller V, Rau T, Schindlbeck C, Rack B, Janni W, Coith C, Beck K, Jänicke F, et al. 2007. Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: a validation study of the CellSearch system. *Clin Cancer Res* **13:** 920–928. doi:10.1158/1078-0432.CCR-06-1695

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43:** e47. doi:10.1093/nar/gkv007

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

RStudio Team. 2022. *RStudio: integrated development environment for R*. RStudio, PBC, Boston. http://www.rstudio.com/.

Salahshourifar I, Vincent-Chong VK, Chang H-Y, Ser HL, Ramanathan A, Kallarakkal TG, Rahman ZAA, Ismail SM, Prepageran N, Mustafa WMW, et al. 2015. Downregulation of *CRNN* gene and genomic instability at 1q21.3 in oral squamous cell carcinoma. *Clin Oral Investig* **19:** 2273–2283. doi:10.1007/s00784-015-1467-7

Sarioglu AF, Aceto N, Kojic N, Donaldson MC, Zeinali M, Hamza B, Engstrom A, Zhu H, Sundaresan TK, Miyamoto DT, et al. 2015. A microfluidic device for label-free, physical capture of circulating tumor cell clusters. *Nat Methods* **12:** 685–691. doi:10.1038/nmeth.3404

Sheikhpour E, Noorbakhsh P, Foroughi E, Farahnak S, Nasiri R, Neamatzadeh H. 2018. A survey on the role of interleukin-10 in breast cancer: a narrative. *Rep Biochem Mol Biol* **7:** 30–37.

Shenoy AK, Lu J. 2016. Cancer cells remodel themselves and vasculature to overcome the endothelial barrier. *Cancer Lett* **380:** 534–544. doi:10.1016/j.canlet.2014.10.031

Siegel RL, Miller KD, Jemal A. 2015. Cancer statistics, 2015. *CA Cancer J Clin* **65:** 5–29. doi:10.3322/caac.21254

Singhal V, Majumdar A. 2018. Majorization minimization technique for optimally solving deep dictionary learning. *Neural Process Lett* **47:** 799–814. doi:10.1007/s11063-017-9603-9

Sinha D, Kumar A, Kumar H, Bandyopadhyay S, Sengupta D. 2018. dropClust: efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Res* **46:** e36. doi:10.1093/nar/gky007

Sinha D, Sinha P, Saha R, Bandyopadhyay S, Sengupta D. 2019. Improved *dropClust* R package with integrative analysis support for scRNA-seq data. *Bioinformatics* **36:** btz823. doi:10.1093/bioinformatics/btz823

Stott SL, Hsu C-H, Tsukrov DI, Yu M, Miyamoto DT, Waltman BA, Rothenberg SM, Shah AM, Smas ME, Korir GK, et al. 2010. Isolation of circulating tumor cells using a microvortex-generating herringbone-chip. *Proc Natl Acad Sci* **107:** 18392–18397. doi:10.1073/pnas.1012539107

Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RM Jr. 1949. *The American soldier: adjustment during army life. (Studies in social psychology in World War II),* Vol. 1, 1: p. 599. Princeton University Press, Princeton, NJ.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102:** 15545–15550. doi:10.1073/pnas.0506580102

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81. doi:10.1038/nature15394

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **71:** 209–249. doi:10.3322/caac.21660

Szczerba BM, Castro-Giner F, Vetter M, Krol I, Gkountela S, Landin J, Scheidmann MC, Donato C, Scherrer R, Singer J, et al. 2019. Neutrophils escort circulating tumour cells to enable cell cycle progression. *Nature* **566:** 553–557. doi:10.1038/s41586-019-0915-y

Tang H, Liu H, Xiao W, Sebe N. 2021. When dictionary learning meets deep learning: deep dictionary learning and coding network for image recognition with limited data. *IEEE Trans Neural Netw Learn Syst* **32:** 2129–2141. doi:10.1109/TNNLS.2020.2997289

Tariyal S, Majumdar A, Singh R, Vatsa M. 2016. Deep dictionary learning. *IEEE Access* **4:** 10096–10109. doi:10.1109/ACCESS.2016.2611583

Thery L, Meddis A, Cabel L, Proudhon C, Latouche A, Pierga J-Y, Bidard F-C. 2019. Circulating tumor cells in early breast cancer. *JNCI Cancer Spectr* **3:** kz026. doi:10.1093/jncics/pkz026

Tian L, Dong X, Freytag S, Lê Cao K-A, Su S, JalalAbadi A, Amann-Zalcenstein D, Weber TS, Seidi A, Jabbari JS, et al. 2019. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat Methods* **16:** 479–487. doi:10.1038/s41592-019-0425-8

Tickle T, Tirosh I, Georgescu C, Brown M, Haas B. 2019. *inferCNV of the Trinity CTAT Project*. Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA. https://github.com/broadinstitute/inferCNV.

Ting DT, Wittner BS, Ligorio M, Vincent Jordan N, Shah AM, Miyamoto DT, Aceto N, Bersani F, Brannigan BW, Xega K, et al. 2014. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep* **8:** 1905–1918. doi:10.1016/j.celrep.2014.08.029

Tirosh I, Izar B, Prakadan SM, Wadsworth MH II, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al. 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352:** 189–196. doi:10.1126/science.aad0501

Tošić I, Frossard P. 2011. Dictionary learning. *IEEE Signal Process Mag* **28:** 27–38. doi:10.1109/MSP.2010.939537

Tsai W-S, Chen J-S, Shao H-J, Wu J-C, Lai J-M, Lu S-H, Hung T-F, Chiu Y-C, You J-F, Hsieh P-S, et al. 2016. Circulating tumor cell count correlates with colorectal neoplasm progression and is a prognostic marker for distant metastasis in non-metastatic patients. *Sci Rep* **6:** 24517. doi:10.1038/srep24517

Turner N, Lambros MB, Horlings HM, Pearson A, Sharpe R, Natrajan R, Geyer FC, van Kouwenhove M, Kreike B, Mackay A, et al. 2010. Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. *Oncogene* **29:** 2013–2023. doi:10.1038/onc.2009.489

Urrutia E, Chen H, Zhou Z, Zhang NR, Jiang Y. 2018. Integrative pipeline for profiling DNA copy number and inferring tumor phylogeny. *Bioinformatics* **34:** 2126–2128. doi:10.1093/bioinformatics/bty057

Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, Hennig BP, Hirche C, Lutz C, Buss EC, Nowak D, et al. 2017. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol* **19:** 271–281. doi:10.1038/ncb3493

Wang L, Balasubramanian P, Chen AP, Kummar S, Evrard YA, Kinders RJ. 2016. Promise and limits of the CellSearch platform for evaluating pharmacodynamics in circulating tumor cells. *Semin Oncol* **43:** 464–475. doi:10.1053/j.seminoncol.2016.06.004

Wang S, Zheng Y, Li J, Yu Y, Zhang W, Song M, Liu Z, Min Z, Hu H, Jing Y, et al. 2020. Single-cell transcriptomic atlas of primate ovarian aging. *Cell* **180:** 585–600.e19. doi:10.1016/j.cell.2020.01.009

Ward MP, Kane LE, Norris LA, Mohamed BM, Kelly T, Bates M, Clarke A, Brady N, Martin CM, Brooks RD, et al. 2021. Platelets, immune cells and the coagulation cascade; friend or foe of the circulating tumour cell? *Mol Cancer* **20:** 59. doi:10.1186/s12943-021-01347-1

Warkiani ME, Guan G, Luan KB, Lee WC, Bhagat AAS, Chaudhuri PK, Tan DS-W, Lim WT, Lee SC, Chen PCY, et al. 2014. Slanted spiral microfluidics for the ultra-fast, label-free isolation of circulating tumor cells. *Lab Chip* **14:** 128–137. doi:10.1039/C3LC50617G

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19:** 15. doi:10.1186/s13059-017-1382-0

Xie J, Girshick R, Farhadi A. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of The 33rd International Conference on Machine Learning* (ed. Balcan MF, Weinberger KQ), Vol. 48 of *Proceedings of Machine Learning Research*, pp. 478–487. PMLR, New York.

Xu L, Mao X, Imrali A, Syed F, Mutsvangwa K, Berney D, Cathcart P, Hines J, Shamash J, Lu Y-J. 2015. Optimization and evaluation of a novel size based circulating tumor cell isolation system. *PLoS One* **10:** e0138032. doi:10.1371/journal.pone.0138032

Yang Y-M, Liu T-H, Chen Y-J, Jiang W-J, Qian J-M, Lu X, Gao J, Wu S-F, Sang X-T, Chen J. 2005. Chromosome 1q loss of heterozygosity frequently occurs in sporadic insulinomas and is associated with tumor malignancy. *Int J Cancer* **117:** 234–240. doi:10.1002/ijc.21175

Yang B, Fu X, Sidiropoulos ND, Hong M. 2017. Towards K-means-friendly spaces: simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning* (ed. Precup D, Teh YW), Vol. 70 of *Proceedings of Machine Learning Research*, pp. 3861–3870. PMLR, International Convention Centre, Sydney, Australia.

Yang X, Deng C, Zheng F, Yan J, Liu W. 2019. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2019, June 16th–June 20th, Long Beach, CA, pp. 4066–4075.

Yip SH, Wang P, Kocher J-PA, Sham PC, Wang J. 2017. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res* **45:** e179. doi:10.1093/nar/gkx828

Yu M, Bardia A, Aceto N, Bersani F, Madden MW, Donaldson MC, Desai R, Zhu H, Comaills V, Zheng Z, et al. 2014. Cancer therapy: ex vivo culture of circulating breast tumor cells for individualized testing of drug susceptibility. *Science* **345:** 216–220. doi:10.1126/science.1253533

Zhang H, Lin X, Huang Y, Wang M, Cen C, Tang S, Dique MR, Cai L, Luis MA, Smollar J, et al. 2021. Detection methods and clinical applications of circulating tumor cells in breast cancer. *Front Oncol* **11:** 652253. doi:10.3389/fonc.2021.652253

Zhao W, Li X, Wang J, Wang C, Jia Y, Yuan S, Huang Y, Shi Y, Tong Z. 2017. Decreasing eukaryotic initiation factor 3C (EIF3C) suppresses proliferation and stimulates apoptosis in breast cancer cell lines through mammalian target of rapamycin (mTOR) pathway. *Med Sci Monit* **23:** 4182–4191. doi:10.12659/msm.906389

Zheng Y, Miyamoto DT, Wittner BS, Sullivan JP, Aceto N, Jordan NV, Yu M, Karabacak NM, Comaills V, Morris R, et al. 2017. Expression of β-globin by cancer cells promotes cell survival during blood-borne dissemination. *Nat Commun* **8:** 14344. doi:10.1038/ncomms14344

Zhou Y, Yang D, Yang Q, Lv X, Huang W, Zhou Z, Wang Y, Zhang Z, Yuan T, Ding X, et al. 2020. Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nat Commun* **11:** 6322. doi:10.1038/s41467-020-20059-6

Received January 16, 2022; accepted in revised form November 10, 2022.