Circulating tumor cell characterization of lung cancer brain metastasis in the cerebrospinal fluid through single-cell transcriptome analysis

Haoyu Ruan[1*], Yihang Zhou[2,3*], Jie Shen[4*], Yue Zhai[2], Ying Xu[2], Linyu Pi[2], RuoFan Huang[5], Kun Chen[6], Xiangyu Li[6], Weizhe Ma[7], Zhiyuan Wu[6], Xuan Deng[1], Xu Wang[3,8,9†], Chao Zhang[2†] and Ming Guan[1†]

[1]*Department of Clinical Laboratory, Huashan Hospital, Fudan University, Shanghai, China*

[2]*Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Shanghai Key Laboratory of Signaling and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai, China*

[3]*Department of Pathobiology, Auburn University, Auburn AL 36849*

[4]*10K Genomics Technology Co., Ltd., 333 Guiping Road, Shanghai, 200233, China*

[5]*Department of Oncology, Huashan Hospital, Fudan University, Shanghai, China.*

[6]*Department of Clinical Laboratory, Huashan Hospital North, Fudan University, Shanghai, China*

[7]*Central Laboratory, Huashan Hospital, Fudan University, Shanghai, China*

[8]*HudsonHudsonAlpha Institute for Biotechnology, Huntsville, AL 35806*

[9]*Alabama Agricultural Experiment Station, Auburn University, Auburn, AL 36849*

*These authors contributed equally to this work.

[†]co-corrresponding authors:

Xu Wang

Phone: +1 (334) 844-7511

Fax: +1 (334) 844-2618

E-mail: xzw0070@auburn.edu

ORCID: 0000-0002-7594-5004

Chao Zhang

Phone: +86 (21) 65986686

Fax: +86 (21) 65986686

E-mail: zhangchao@tongji.edu.cn

Ming Guan

Phone: +86 (21) 52889999

Fax: +86 (21) 62481061

E-mail: guanming88@yahoo.com

Key words: lung adenocarcinoma, Leptomeningeal metastases, cerebrospinal fluid, circulating tumor cell, single cell RNA sequencing

Abstract

Metastatic lung cancer accounts for about half of the brain metastases (BM). Development of leptomeningeal metastases (LM) are becoming increasingly common, and its prognosis is still poor despite the advances in systemic and local approaches. Until now, the cytology analysis in the cerebrospinal fluid (CSF) remains the gold diagnostic standard. Although several previous studies performed in CSF have offered great promise for the diagnostics and therapeutics of LM, a comprehensive characterization of CTCs in CSF is still lacking. To fill this critical gap of lung adenocarcinoma leptomeningeal metastases (LUAD-LM), we analyzed the transcriptomes of 1,375 cells from 5 LUAD-LM patient and 3 control samples using single-cell RNA sequencing technology. We defined CSF-CTCs based on abundant expression of epithelial markers and genes with lung origin, as well as the enrichment of metabolic pathway and cell adhesion molecules, which are crucial for the survival and metastases of tumor cells. Elevated expression of *CEACAM6* and *SCGB3A2* was discovered in CSF-CTCs, which could serve as candidate biomarkers of LUAD-LM. We revealed substantial heterogeneity in CSF-CTCs among LUAD-LM patients and within patient among individual cells. Cell-cycle gene expression profiles and the proportion of CTCs displaying mesenchymal and cancer stem cell properties also vary among patients. In addition, CSF-CTC transcriptome profiling identified one LM case as cancer of unknown primary site (CUP). Our results will shed light on the mechanism of LUAD-LM and provide a new direction of diagnostic test of LUAD-LM and CUP cases from CSF samples.

Introduction

Lung cancer is the second most common cancer type in both men and women (Siegel et al., 2016). Non-small cell lung cancer (NSCLC) is the main type of lung cancer, accounting for 85% of lung malignancies with a 5-year survival rate less than 15% (Chen et al., 2016). Histologically, NSCLC was further clarified into three subtypes: lung adenocarcinoma (LUAD), squamous-cell carcinoma and large cell carcinoma (Herbst et al., 2008), among which LUAD is the most common histological subtype (Youlden et al., 2008). Most lung cancer-associated morbidity results from metastases. Brain is the most common metastatic site of NSCLC and the incidence of brain metastases (BM) is ranging from 22% to 54%, occurring at different stages of tumorigenesis, especially in advanced patients (Alberg et al., 2005; Grinberg-Rashi et al., 2009). Of all cancer patients with brain metastases, lung cancer is the primary tumor in 40 to 50% cases (Schouten et al., 2002), which is the highest among all cancer types and equals all other primary cancer types combined (Barnholtz-Sloan et al., 2004). The lethality of brain metastasis follows severe deterioration of patients' quality of life, and symptoms can include headaches, seizures, cognitive/motor dysfunction and coma (Lassman and DeAngelis, 2003; Peters et al., 2016; Sankey et al., 2019). Leptomeningeal metastasis (LM) occurs in 3-5% of patients with advanced NSCLC and it is the most frequent in the LUAD subtype (Remon et al., 2017). Leptomeningeal metastases results from dissemination of cancer cells to both the leptomeninges (pia and arachnoid) and cerebrospinal fluid (CSF) compartment (Gleissner and Chamberlain, 2006; Grossman and Krabak, 1999). In this study, we enrolled five LUAD patients of leptomeningeal metastases (LUAD-LM), four patients showed leptomeningeal enhancement by magnetic resonance imaging (MRI) of brain, and one patient only detected tumor cells in CSF.

The diagnosis and monitoring of NSCLC-LM are based on neurological and radiographic imaging as well as CSF examinations. Until now, a positive CSF cytology result remains the gold standard for the diagnosis of NSCLC-LM and the gadolinium-enhanced MRI of the brain and spine is the best imaging technique (Cheng and Perez-Soler, 2018). In recent years, CSF liquid biopsy has attracted more and more attention. Tumor cells in CSF were considered as circulating tumor cells (CTCs) as CTCs in blood (Lin et al., 2017). The CellSearch technique (Tu et al., 2015), which utilizes immunomagnetic selection, identification and quantification of CTCs in the CSF, is more sensitive than the conventional cytology and MRI for the diagnosis of LM (Jiang et al., 2017; Nayak et al., 2013; Tu et al., 2015). Moreover, evaluation of CSF circulating free DNA (cfDNA) has the potential to facilitate and supplement LM diagnosis. The detection of genomic alterations can also inform targeted therapy (De Mattos-Arruda et al., 2015; Sasaki et al., 2016). The aim of NSCLC-LM treatment is palliative including corticosteroids, chemotherapy, and radiotherapy which can extend survival to an average of less than one year (Lowery and Yu, 2017). Intrathecal chemotherapy, molecularly targeted therapy and immunotherapy are also effective treatment to prolong survival (Cheng and Perez-Soler, 2018). The development of new technologies and approaches improved the diagnosis and therapy of NSCLC-LM, but the outcomes of brain metastases are still poor. To make breakthroughs in tackling the clinical challenge of brain metastasis, a comprehensive understanding of its mechanisms is needed. CSF sampling through lumbar puncture can help diagnose and monitor the progress of brain metastases, with the advantage of minimal risk compared to brain biopsy. By analyzing the transcriptome characteristics of tumor cells in CSF of NSCLC-LM patients, new biomarkers could be discovered for clinical diagnosis. In addition, the knowledge on gene

expression alterations in brain metastases compared to primary tumors will help develop novel targeted therapeutic approaches with improved efficacy.

CSF-CTCs was proposed as a novel tool for the diagnosis of LM (Lin et al., 2017). However, tumor cells are relatively rare in patient CSF samples, and ≥1 CSF-CTC/mL was defined as a cutoff for diagnosis (Lin et al., 2017). The traditional profiling technologies that measure tumor cells in bulk was confounded by the normal lymphocytes and cannot capture the gene expression heterogeneity among tumor cells. Therefore, we investigated the transcriptome characteristics of CSF-CTCs by Smart-seq2 single-cell RNA sequencing (scRNA-seq). The CSF-CTCs have not been investigated at the single-cell transcriptome level in leptomeningeal metastases. To date, there is only one CSF scRNA-seq study in the literature, which identified rare CNS immune cell subsets that may perpetuate neuronal injury during HIV infection (Farhadian et al., 2018). Here, we investigate the transcriptional profiles for more than one thousand CSF-CTCs from five LM patients, defined the CSF-CTC transcriptome profile and revealed intra-tumoral and inter-tumoral heterogeneity of patients for the first time.

Results

Cell composition of the cerebrospinal fluid at single-cell transcriptome level.

To characterize the single-cell transcriptomes and the composition of CSF cells under healthy conditions, three samples (N1-N3) were used from patients who were screened for potential brain infections and diagnosed as normal (Table S1). We sequenced the single-cell transcriptomes of 576 CSF cells using Smart-seq2 single-cell RNA-seq (scRNA-seq) technology (Figure 1A). Besides CSF cells, blood T cells and B cells were flow sorted and sequenced to establish the cell type transcriptome profiles to define the normal CSF cells composition. After quality filtering (see Methods), 207 normal CSF cells, 41 B cells and 41 T cells are clustered using t-distribution stochastic neighbor embedding (t-SNE) method (Figure 1B). On average, 803 expressed genes were detected per individual cell. We identified three clusters corresponding to B cells, T cells and monocytes defined by the expression patterns of lymphocyte markers (Figure 1C and Figure S2). The Blood-T and CSF-T cells are clustered together, indicating normal lymphocytes have similar expression profiles in different microenvironments. Normal CSF samples consist of 80.3% T cells and 19.7% monocytes. No B cells were found in CSF samples (Table S2). Our result is consistent with previously known proportions in the textbook (Höftberger et al., 2015), which serves as a proof of principle of scRNA-seq in CSF samples.

Identification and characterization of circulating tumor cells in the CSF of LUAD-LM patients.

In order to determine whether CTCs could be isolated from the CSF samples of LUAD-LM patients, seven LM patients were enrolled in the scRNA-seq study (Table S1). Candidate

CTCs are defined as CD45– and Calcein Blue AM+ cells with significant larger cell diameters than leukocytes. In total, 1,776 candidate CTCs from five LUAD-LM patients (P1, P2, P4, P6 and P7) were FACS sorted (Figure S1) and sequenced using the Smart-seq2 protocol (Figure 1A). 1,152 cells from patient samples with at least 600 covered genes in their transcriptomes are included in our analysis (Table S3), and these cells are clustered using the t-SNE method along with three normal CSF samples. CSF cells from patient samples were clustered according to patient of origin, and they do not cluster with normal CSF cells (only 15 cells of P4 in monocyte cluster), suggesting that they could be CTCs circulated from the brain metastatic tumor (Figure 2A). To exclude the possibility of technical variability, CSF samples (P1-1 and P1-2) were collected from the same patient (P1) within a two-month time interval, so independent cell sorting, library preparation and sequencing were performed. Cells from P1-1 and P1-2 formed a single coherent cluster, indicating that the individual heterogeneity we observed is not due to technical variation (Figure 2A). Sequencing and mapping quality are other common factors for technical artifact. For the analysis of candidate CTCs, a minimum of 1,000 covered genes per cell cutoff was applied and 967 candidate CTCs remained with an average of 2,070 genes per cell. No significant heterogeneity in mapping quality and gene coverage were detected across patient samples (Figure S3).

After we established the CTC isolation protocols from patient CSF samples, we tested whether cell morphology alone without CD45– selection could achieve similar results. Live cells were isolated from another LUAD-LM patient (P3) and candidate CTCs were selected for scRNA-seq merely based on cell morphology. Cell identity analysis results revealed 100% of collected cells from P3 CSF samples are monocytes (Figure S4). Therefore, we conclude that the CD45 negative selection is necessary for successful CTC isolation from patient CSF samples.

To determine whether the patient CSF cells are actually CTCs, we investigated the single-cell gene-expression patterns and discover that all cells in P1, P2, P6 and P7 CSF samples were CTCs and 90.4% (142/157) of P4 CSF cells are CTCs (Table 1). At molecular level, we define CSF-CTCs as non-immune cells with transcriptome signatures for lung adenocarcinoma markers, epithelial markers and proliferation markers (Figure 2B and Figure S5). Compared to control CSF samples, patient CSF samples lack expression from immune cell markers (Figure 2B) and have a mean ImmuneScore (Yoshihara et al., 2013) of 127, which is significantly lower than the normal CSF cells (mean ImmuneScore = 1,703, $P < 0.001$, Wilcoxon Rank-Sum test; Figure 2C). Epithelial cell markers, including *EPCAM*, *CDH1*, *KRT7*, *KRT8*, *KRT18* and *MUC1*, are expressed on in CTCs, suggesting their epithelial origin (Figure 2B and Figure S5). Lung adenocarcinoma markers *SFTPA1*, *SFTPA2*, *SFTPB* and *NAPSA* are expressed in CTCs, indicating that they are originated from tumor cells of primary lung cancer (Figure 2B and Figure S5). CTCs also have expression of common proliferation markers *CCND1* and *TOP2A* (Figure 2B and Figure S5).

Transcriptome signatures of CSF-CTCs in LUAD-LM patients.

Among the significantly expressed genes between patient and normal CSF cells, *CEACAM6*, *SCGB3A2*, *MMP7* and *C3* were highly upregulated in CTCs (Figure 2D). *CEACAM6* (adjusted *P*-value = 1.23E-76; $\log_2$ fold-change = 6.22) is carcinoembryonic antigen cell adhesion molecule, which is a biomarker for mucinous adenocarcinoma including colon, pancreas, breast, ovary and lung malignancies. Overexpression of *CEACAM6* has been shown to associate with poor prognosis due to its roles in cellular invasiveness, resistance to anoikis and metastatic potential (Beauchemin and Arabzadeh, 2013). Another marker for pulmonary carcinoma (Kurotani et al., 2011), *SCGB3A2*, is also significantly upregulated in our dataset

(adjusted *P*-value = 4.46E-33; $\log_2$ fold-change = 3.06). *SCGB3A2* is a member of the secretoglobin (SCGB) gene superfamily and mainly found in bronchial epithelial cells. It is a growth factor during fetal lung development (Kurotani et al., 2008) with anti-inflammatory function in the lung (Chiba et al., 2006). As secreted proteins, elevated expression of *CEACAM6* and *SCGB3A2* have great potential in developing CSF immunoassay for LUAD-BM diagnosis.

MMP7 (adjusted *P*-value = 2.94E-49; $\log_2$ fold-change = 3.99), matrix metalloproteinase 7, is exclusively expressed in the epithelial cells promoting carcinogenesis through induction of epithelial-to-mesenchymal transition (EMT) (Wilson and Matrisian, 1996). Cancer cells can secrete MMPs to disrupt the basement membrane (BM), facilitating the metastases to other sites (Sevenich and Joyce, 2014). MMPs can also open the blood-brain barrier by degrading tight junctions proteins (Yang and Rosenberg, 2011). C3 (adjusted *P*-value = 8.13E-33; $\log_2$ fold-change = 2.19) derived from lung cancer cells in CSF activates the C3a receptor in the choroid plexus epithelium to disrupt the blood-CSF barrier which allows plasma components, including amphiregulin, and other mitogens to enter the CSF and promote cancer cell growth (Boire et al., 2017). The gene expression profiles of candidate CTCs are in sharp contrast to the CSF lymphocytes, and these transcriptome characteristics are reassuring that the patient CSF cells are indeed CSF-CTCs.

To determine the functional enrichment in CSF-CTC transcriptomes, we performed gene set enrichment analysis (GSEA) and discovered two major categories were significantly enriched at an FDR (false discovery rate) cutoff of 0.05: energy metabolism pathways and cell adhesion pathways (Figure 2E). Energy metabolism pathways, including glycolysis gluconeogenesis pathway and pentose phosphate pathway, are critical for tumor growth and the energy demand in the brain. The up-regulated cell adhesion pathways consist of tight junction, ECM (extracellular

matrix) receptor interaction, adherens junction, and focal adhesion, indicating that CSF-CTCs possessed a higher adhesion strength which is crucial for a number of essential functions such as survival, proliferation and migration, providing the ability to maneuver through the capillary-sized vessels to a new location (Craig and Brady-Kalnay, 2011; Mao et al., 2018). In addition, several tumor-related signaling pathway (TGF beta signaling pathway, WNT signaling pathway, P53 signaling pathway) were enriched in CSF-CTCs at a relaxed FDR cutoff of 0.20 (Figure 2E).

Spacial and gene expression heterogeneities of LUAD-LM tumors.

LM can occur at different brain locations, resulting in spacial heterogeneity of the metastatic tumors. We examined the five LUAD-LM patients and this is exactly the case (Figure 3A). To determine the single-cell gene expression heterogeneities among patients, we quantified the pairwise correlations between the expression profiles of 967 single-CTC transcriptomes from the five LUAD-LM samples (Figure 3B) and discovered significant expression heterogeneity both among different patient CSF samples and within individual patient CSF sample (correlation coefficients ranging from –0.057 to 0.829). Inter-tumor heterogeneity between different LUAD-LM patients is significantly greater than the intra-tumoral heterogeneity within individual LUAD-LM patient (mean correlation coefficient –0.009 vs. 0.029, *P*-value < 2.2e-16, Wilcoxon Rank-Sum test, Figure 3C). To compare the single-cell correlation profile with the primary tumors, we utilized single-cell expression data (GSE69405) from human non-small cell lung cancer cell line H358, and human lung adenocarcinoma patient-derived xenograft (PDX) MBT15 and PT45 (Kim et al., 2015). The cell-to-cell correlations within individual primary tumor samples are significantly higher than that within individual patient CSF samples (mean

correlation coefficient 0.092 vs. 0.029, *P*-value < 2.2e-16, Wilcoxon Rank-Sum test), indicating greater heterogeneity in clinic CSF samples (Figure 3B-C).

To identify patient-specific signatures, we profiled the differentially expressed genes (DEGs) within each of the give patients and more than 600 DEGs were identified in individual CSF samples at a fold change cutoff of 1.5 (*P*-value < 0.001, Figure 3D). The important gene families that vary across patients include serpins (SERPIN), galectins (LGALS), claudins (CLDN), secretoglobins (SCGB) and solute carrier family (SLC). These gene family members are the major contributors of the heterogeneity among patients.

The majority of the CSF-CTCs are in the non-cycling state in LUAD-LM patients

LUAD-LM patients tend to have poor prognosis. After treatment, residual tumor cells disseminate rapidly in CSF within several months. We analyzed the cell-cycle state of the CSF-CTCs based on the single-cell transcriptomes. On average, high-cycling cells only account for 7.2% in LUAD-LM patient (4% in P1, 11% in P2, 14% in P4, 3% in P6 and 4% in P7), which is much fewer than the H358 cell line (36%) and two PDX samples (33% in MBT15, 25% in PT45) (Figure 4A and Figure S6).

Cancer stemness and partial epithelial-to-mesenchymal transition (EMT) in CSF-CTCs.

Cancer stem cells (CSC) has the properties of asymmetrically dividing, differentiation and self-renew, as well as increased intrinsic resistance to therapy (MacDonagh et al., 2016). The expression levels of lung CSC candidate biomarkers were investigated, including *PROM1*(*CD133*) (Bertolini et al., 2009), *CD44* (Leung et al., 2010), *ALDH1A1* (Sullivan et al., 2010), *ALDH1A3* (Shao et al., 2014), *ALDH3A1* (Patel et al., 2008) and *ABCG2* (Bleau et al., 2009) (Figure S7). Elevated aldehyde dehydrogenase (ALDH) activity is a hallmark of CSC in

NSCLC. Among the three ALDHs we examined, 162 (16.7%) CSF-CTCs have detectable expression of *ALDH1A1*, with much fewer cells expressing *ALDH1A3* (21 cells) or *ALDH3A1* (46 cells). 27.7% (268/967) CSF-CTCs across five LUAD-LM patients have CD44 expression, which is a stem cell marker for CTC aggregation and polyclonal metastases (Liu et al., 2019). *PROM1* and *ABCG2* positive CSF-CTCs are extremely rare (Figure S7). 57 CTCs have both expression of CD44 and ALDH1A1 and no CSF cells have all CSC makers positive.

Epithelial-to-mesenchymal transition (EMT) has been suggested as a driver of epithelial tumor spreading (Lambert et al., 2017). During the EMT process, epithelial cells lose cell-cell adhesion and cell polarity, in order to gain migration and invasion capabilities to behave like multipotent mesenchymal stem cells. Almost all CSF-CTCs have high expression of epithelial markers (Figure 4B-C and Figure S8). However, we discovered partial EMT process in these CSF-CTCs, which is defined as tumors cells exhibiting both mesenchymal and epithelial characteristics (Saitoh, 2018). Three markers (*FN1*, *VIM*, *CD44*) were analyzed to calculate the Gene Set Variation Analysis (GSVA) score of mesenchymal/CSC hybrid phenotypes (Hanzelmann et al., 2013). 113 cells in P1 (33.2%) and 54 cells in P4 (42.5%) have GSVA scores of epithelial and mesenchymal/CSC markers are both greater than 0.5 (Figure 4B), suggesting partial EMT in these patients. However, other patients only have a few cells showing both epithelial and mesenchymal/CSC characteristics (6/122 in P2, 1/206 in P6, 15/172 in P7) (Figure S8). Although these CSF-CTCs have features of classical EMT, they lack expression of N-cadherin, classical EMT transcription factors (*ZEB1/2*, *TWIST1/2* and *SNAIL1/2*) or the EMT regulator TGFβ (Puram et al., 2017).

We also examined the extracellular matrix (ECM) related markers, which is another class of EMT features. Compared to normal CSF cells, the ECM receptor interaction KEGG pathway

was significantly enriched in CSF-CTCs (FDR<0.05, Figure 2E). We selected core enrichment genes of ECM receptor interaction pathway, including laminins (*LAMA3*, *LAMA5*, *LAMB2*, *LAMC1*) (Theocharis et al., 2016), integrins (*ITGA3*, *ITGB4*) (Giancotti and Ruoslahti, 1999) and *CD47* (Sick et al., 2012). Abundant expression of ECM genes is observed in all patients, which could be a common feature of CSF-CTCs (Figure 4B-C; Figure S8). These results suggest that the upregulation of ECM related genes may contribute to the generation of CTCs from solid tumor sites or to the survival of cancer cells deprived of microenvironmental signals as they circulate in the CSF.

Cancer-testis antigens (CTAs) in CSF-CTCs contribute to the among patient heterogeneity

Tumor cells frequently express cancer-testis antigens (CTAs) whose expression is typically restricted to germ cells, providing unprecedented opportunities for clinical development of cancer diagnosis and immunotherapy (Salmaninejad et al., 2016). A recent study has demonstrated the extensive heterogeneity of CTAs in LUAD single-cell data (PDXs and cell lines) (Ma et al., 2019). However, little is known about the heterogeneity of all possible CTAs expressed in CSF-CTCs from LUAD origin. We examined the expression of 276 selected CTAs (http://www.cta.lncc.br/modelo.php) in CSF-CTCs and discovered substantial inter-tumor heterogeneity and intra-tumor heterogeneity (Figure 4D; Figure S9). Expression of *XAGE1B* is only observed in P1, P6 and P7, whereas *BRDT* expression is restricted in P6 and P7 (Figure 4D). LY6K is specific to subset of CTCs in P4 and P6. Overall, patients P1 and P4 have significantly larger numbers of expressed CTAs than other patients (Figure 4E). *SPAG9* is ubiquitously expressed in 41.3% of CTCs in all five patients (399/967) at high level, with the potential to serve as a target for immunotherapy (Figure S9) (Wang et al., 2013).

Characterization of a case of cancer of unknown primary site (CUP) through CSF-CTC single-cell transcriptomes.

Patient P8, a 49-year-old male, was diagnosed with cancer of unknown primary site (CUP) in 2017. CUP is a well-recognized clinical disorder, accounting for 3-5% of all malignant epithelial tumors. Metastatic adenocarcinoma is the most common CUP histopathology (80%) (Pavlidis and Pentheroudakis, 2012). P8 showed multiple metastases including cervical lymph nodes and LM (Figure S10A). The immunohistochemistry (IHC) results of left lymph nodes biopsy indicated that the metastatic adenocarcinoma with positive epithelial markers (CK pan, CK7, CK8, CK18, CK19 and MUC1; Figure S10B-C) and a prolactin-induced protein PIP/GCDFP15 (Figure S10E), which is a small secreted glycoprotein whose expression is generally restricted to cells with apocrine properties (Urbaniak et al., 2018). The proliferation marker MKI67 is partially positive (Figure S10F). Therefore, the primary tumor is epithelial origin with apocrine properties. Based on the IHC results, we could exclude the possibility of the following locations of the primary tumor: lung cancer (markers NAPSA–, TTF1–, P63–, synaptophysin/SYP–; Figure S10D) (Mengoli et al., 2018; Yatabe et al., 2019), gastrointestinal cancer (VIL1/Villin–; Figure S10G) (Bacchi and Gown, 1991), prostate cancer (KLK3/PAS–; Figure S10H) (Paliouras et al., 2007) and liver cancer (GPC3–; Figure S10I) (Filmus and Capurro, 2013). P8 had partial response (PR) to chemotherapy (paclitaxel liposome for injection and cisplatin), but the disease had recurred in May 2019 with leptomeningeal metastases after the treatment (Figure S10A). We performed single-cell RNA sequencing of CSF-CTCs to help diagnose primary site of P8.

P8 CSF-CTCs form a single cluster on the t-SNE clustering plots, independent from LUAD-LM CTCs and normal CSF cells (Figure 5A). There is some degree of separation of samples P8-1 and P8-2, which were collected with a six-month time interval, reflecting potential disease progression (Figure 5A). P8 CSF-CTCs were defined by the epithelial signature and lack of CD45 expression (Figure 5B). None of the lung origin markers are expressed, which is in sharp contrast to the LUAD-LM patients (Figure 5B) and is consistent with the immunohistochemistry results (Figure S10D). In addition, upregulated genes in LUAD-LM CTCs (*MMP7*, *SCGB3A2*, *C3*, *CDH1* and *EGFR*) are not DEGs in P8 CTCs, except for *CEACAM6* which is shared across P8 and all LUAD-LM patients (Figure S11A-B). The gene set enrichment analysis (GSEA) revealed active metabolism and tight junction pathway (FDR<0.05) with a few other KEGG pathways as the characteristics of P8 CTCs (Figure S11C).

Based on the single-cell transcriptomes, we found PIP gene is specifically expressed in the P8 CTC cluster at a high level (Figure 5B and Figure S11A-B), which is consistent with immunohistochemistry results of left lymph nodes biopsy tissue (Figure S10D). PIP is a cytoplasmic marker commonly used to identify breast cancer, but not exclusively, as its expression was also found in several other types of human cancers including prostate, sweat and salivary gland cancer (Urbaniak et al., 2018). *ANKRD30A* (NY-BR-1) is considered as a breast differentiation antigen that could represent a suitable target for immunotherapy of breast cancer (Seil et al., 2007; Theurillat et al., 2007; Wang et al., 2006). The expression of *ANKRD30A* is restricted to normal breast, normal testis, normal prostate and also detected in breast cancer as a breast (cancer) specificity marker (Lacroix, 2006; Varga et al., 2006) and in prostate cancer (RNA level) (Jager et al., 2005). It is interesting that 7 CTCs of P8 have high expression of the *SCGB2A2* genes (Figure 5C), a carcinoma marker of breast origin including primary tissues,

metastatic tissues and blood-CTCs (Julian et al., 2008; Markou et al., 2011; Reinholz et al., 2011; Tafreshi et al., 2011; Zehentner and Carter, 2004). *SCGB2A2* is also positive in some tissues of gynecologic malignancies (Zafrakas et al., 2006), but P8 is a male. In addition, *SCGB2A2* is also associated with salivary gland cancer (Forner et al., 2018).

P8 CSF-CTCs with more than 1000 genes covered were analyzed to investigate tumor transcriptome characteristics. The proliferation ability of P8 CTCs is similar to LUAD-LM CSF-CTCs, with 3/131 (P8-1) and 11/262 (P8-2) cells in high-cycling state (Figure 5D). Interestingly, a large proportion of (118/393) P8 cells have *PROM1* (*CD133*, a classic CSC marker) expression, which is different from LUAD CSF-CTCs (20/967). The expression levels of epithelial, mesenchymal and CSC genes were also examined. All P8 CTCs have high expression of epithelial genes. CTCs (290/393) with stem-like phenotype (PROM1+CD44) are in epithelial state (Miyamoto et al., 2016) and about 10% (36/393) cells have detectable expression of mesenchymal genes (FN1+VIM) (Figure 5E).

Methods

Patients information and sample collection.

All human sample materials used in this research are collected at Huashan Hospital, Fudan University. The consent forms and the proposed studies were approved by Ethics Committee. Clinical information of patients was listed in Table S1.

Cell Sorting and single-cell preparation

Antibody (CD45, CD3, CD19, BD Biosciences) and labeling dye for live cells (Calcein Blue AM, Life Technologies, CA) were used per manufacturer recommendations. Live cells (Calcein Blue AM+) in normal CSF samples (N: N1-N3) and live tumor cells (Calcein Blue AM+, CD45–) in patient CSF samples (P: P1-P8) were selected for sequencing (Figure 1A, Table S1). Among all of the CSF samples, 3,792 single cells were selected for sequencing (N: 624 cells; P: 3,168 cells; Table S1). 168 blood-T cells (Calcein Blue AM+, CD45+, CD3+) and 168 blood-B cells (Calcein Blue AM+, CD45+, CD19+) were also sorted for sequencing (Table S1).

Target cells were sorted into pre-prepared 96-well plates by FACS (fluorescence-activated cell sorting). Single-cell lysates were sealed, vortexed, spun down, placed on dry ice and transferred immediately for storage at -80°C.

SMRAT-seq2 library construction and sequencing

Library for isolated single cell was generated by SMART-Seq2 method (Picelli et al., 2013) with the following modifications: RNA were reverse transcribed with Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific, MA), and whole transcriptome were amplification using KAPA HiFi Hot Start Ready Mix (KAPA Biosystems, MA). cDNA library was purified using

Agencourt XP DNA beads (Beckman Coulter, CA) and quantified with a high sensitivity dsDNA Quant Kit (Life Technologies, CA). It is worth mentioning that full length cDNA libraries were tagmented, and then only 3' end sequence (500-1000bp) were amplified and enriched for sequencing on an Illumina HiSeqX machine, which is different from traditional SMART-Seq2 method of full tagmented-libraries sequence.

scRNA-seq expression analysis

The Illumina sequencing data were demultiplexed based on sample barcodes. Adapter sequences, poly T and residue barcodes were trimmed using custom scripts. We used UMI-tools (Smith et al., 2017) to remove UMIs and trim-galore (Kruger, 2012) to remove low-quality bases. The filtered reads were aligned to the human reference genome (hg19) by STAR (Dobin et al., 2013) and BAM files were prepared by SAMtools (Li et al., 2009). Gene expression counts were obtained by FeatureCounts (Liao et al., 2014). Fastqc (Andrews, 2010) and multiqc (Ewels et al., 2016) were used for quality control and reports during each step.

Genes expressed in less than 10 cells were filtered out from gene expression matrix of CSF samples. Individual cells with fewer than 600 covered genes and over 20% mitochondrial reads were filtered out, and 1,986 single cells remained (401 immune cells and 1,585 CTCs) for subsequent analysis using the Seurat 3.0 software package (Butler et al., 2018; Stuart et al., 2019) (Table 1and S2, Figure 1, 2 and 5A-C, Figure S2, 4, 5, 11 and 12).The mean number of genes detected per cell is 830 for immune cells and 1,870 of tumor cells, respectively.

When we analyze the transcriptome characteristics of CTCs, we selected tumor cells with more than 1000 covered genes and have 1360 CTCs retained (340/ P1, 122/P2, 127/P4, 206/P6,

172/P7, 393/P8) for analysis (Figure 3, 4, and 5D-E, Figure S3, S6-S9 and S13). The mean number of genes detected per LUAD CSF-CTCs was 2070.

Clustering and marker expression analysis for cell type identification

Cells were clustered by non-supervised t-SNE dimensionality reduction (van der Maaten and Hinton, 2008) based on their gene expression counts. The cells were separated into groups with indication of cryptic inner-group connection. The cluster-specific marker genes were identified by the FindAllMarkers function in Seurat 3.0. Single-cell RNA-Seq data of two human lung adenocarcinoma patient-derived xenograft (PDX) samples (LC-PT-45, PT45; LC-MBT-15, MBT15) and human non-small cell lung cancer cell line (H358 cell line) were obtained from NCBI Sequence Read Archive with accession number GSE69405 (Kim et al., 2015). These data were filtered using the same pipeline as CSF-CTCs and only cells with more than 1,000 genes were included for analysis.

To infer the cell type identity, Seurat 3.0 was used to generate expression heatmaps of selected gene markers of known cell types, including T cells (*CD2*, *CD3D*, *CD3E* and *CD3G*), B cells (*CD19*, *MS4A1*, *CD79A* and *CD79B*), monocyte (*CD14*, *CD68* and *CD163*), lung cells (*SFTPA1*, *SFTPA2*, *SFTPB* and *NAPSA*), epithelial cells (*EPCAM*, *CDH1*, *KRT7*, *KRT8*, *KRT18* and *MUC1*), proliferative cells (*CCND1* and *TOP2A*) (Glasser et al., 2005; Kim et al., 2014; Miyamoto et al., 2015; Peng et al., 2019; Poornima et al., 2002; Sin et al., 2013; Strayer et al., 2002). The ImmuneScore was computed based on ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumors using Expression data) R package (Yoshihara et al., 2013).

Differential gene expression and pathway enrichment analysis

Significantly differential expression gene (DEGs) between samples were detected by DESeq2 (Love et al., 2014) using normalized gene expression counts, at a adjust p-value cutoff of 0.05 and a fold-change cutoff of 2. GSEA (gene set enrichment analysis) (Subramanian et al., 2005; Zhao et al., 2006) were used for functional enrichment analysis of KEGG pathways (Kanehisa, 2019; Kanehisa and Goto, 2000; Kanehisa et al., 2019; Zhao et al., 2006).

Cell cycle analysis

Cell cycle assignment was performed in R version 3.6.0 using the CellCycleScoring function included Seurat 3.0 package. Cycling cells are defined as cells with either G1/S.Score or G2/M.Score greater than 0.2. The criterion of intermediate cells is 0 < G1/S.Score or G2/M.Score ≤0.2. The rest cells are defined as non-cycling cells.

Discussion

The value of CSF samples in the detection, diagnostics and biomarker characterization in lung cancer brain metastases.

The majority of leptomeningeal metastases (LM) is from solid malignancy of primary breast and lung cancers (Waki et al., 2009) and LM remains a clinical challenge. The cerebrospinal fluid (CSF) is in direct contact with tumor cells in leptomeningeal metastases, therefore liquid biopsy of CSF will reflect the real-time status of LM. CSF cfDNA (cell-free DNA) can be used to characterize and monitor the development of NSCLC-LM, and mutations detected in the cfDNA will inform the clinical therapy (Sasaki et al., 2016), and a recent research revealed 53.2% cfDNA mutations in patients with cancer had features consistent with clonal hematopoiesis which were not real mutations of tumors (Razavi et al., 2019). However, cell-free method has its limitation due to the inability to trace the tumor cells. It has been shown that tumor cell counts in CSF is associated with clinic prognosis of lung cancer LM (Tu et al., 2015), but further studies were needed to fully characterize the CSF-CTCs in LUDA-LM patients. To fill this critical gap, our study serves as the first single-cell transcriptome profiles of LUAD CSF-CTCs by scRNA-seq in multiple patients. We established a protocol for successful isolation of circulating tumor cells for scRNA-seq in CSF samples. The CSF-CTC population is enriched for meta-static precursors, and their single-cell transcriptome signatures will facilitate early detection of LUAD-LM and the identification of potential therapeutic targets.

Single-cell transcriptome signatures defining the CSF-CTCs in lung adenocarcinoma leptomeningeal metastases (LUAD-LM).

To define the LUAD-LM CSF-CTCs, we discovered a number of marker genes by comparing the patient CSF-CTCs with normal CSF cells. Several lung cancer markers and proliferation makers were significantly higher expressed. The expression of cycle gene *CCND1* in CSF-CTCs (596/967) were much broader and higher than MKI67, which a classical proliferation marker expressed in 12.5% (121/967) CSF-CTCs (Figure 2B,D). Therefore, *CCND1* is more suitable to serve as proliferation marker for LUAD CSF-CTCs diagnosis.

TTF1 is a major transcription factor regulating genes expression in airway epithelial cells, including surfactant proteins *SFTPA* (Bruno et al., 1995), *SPTPB* (Bohinski et al., 1994), *SFTPC* (Kelly et al., 1996) and *SCGB1A1* (CC10) (Ray et al., 1996). *TTF1* has been used for clinical diagnosis of human LUAD (Di Loreto et al., 1997; Nakamura et al., 2002). However, its expression in CSF-CTCs was not upregulated compared to immune cells which limits its application in LUAD-CTCs identification.

SLC34A2 is another commonly expressed gene in CSF-CTCs (Figure 2D). Its expression was only found in the apical membrane of type II alveolar epithelium cells (ATII) (Traebert et al., 1999; Wang et al., 2015), providing the potential as a candidate LUAD CSF-CTC specific marker. Secretoglobin gene superfamily are diagnostic markers for cancers, and the most studied member *SCGB1A1* (CC10) is considered as a tumor suppressor and is expressed in less than 10% of human NSCLCs (Linnoila et al., 2000). In contrast, *SCGB3A2* is a useful immunohistochemical marker for NSCLCs, particularly adenocarcinomas (Kurotani et al., 2011). SCGB1A1 is not expressed in CSF-CTCs, whereas SCGB3A2 are broadly and highly expressed in the majority of the cells, which could serve as a novel marker for LUAD CSF-CTCs (Figure 2D).

Energy metabolism and cell adhesion pathways are significantly enriched in LUDA CSF-CTCs.

Glucose is the main energy metabolite and it is transported through endothelial cell glucose transporter GLUT-1 across blood-brain barrier to meet the high energy demands of the brain tissue (Mergenthaler et al., 2013). Previous study using experimental brain metastases models have shown that the utilization of glucose was enhanced in both glycolytic and pentose phosphate pathways (Chen et al., 2007). This is also observed in patient CSF-CTCs. with glucose-associated pathways significantly up-regulated (Figure 2E), which could be important for tumor growth. Tumor cells also have active and rich amino acid metabolism to ensure the sufficient protein synthesis for tumor survival.

The adhesion between CTCs and endothelial cells adhesion is essential in tumor metastasis (Green et al., 2016), by facilitating the circulation and invasion of the metastatic cells (Giladi and Amit, 2017). The up-regulation of adhesion-related genes in CSF-CTCs is also important for CTCs clusters formation, which have been detected by CellSearch technology previously (Tu et al., 2015). Blood-CTCs cluster termed “circulating tumor microemboli” (CTM), or CTC-WBC cluster have been described with higher malignancy and had a better change to survive in bloodstream (Aceto et al., 2014; Szczerba et al., 2019). In addition, plakoglobin (JUN) was identified as a key mediator of tumor cells clustering in human breast CTCs. We detected the expression of plakoglobin in a subset of CSF-CTCs, but the critical gene for CTC-neutrophil cluster formation, *VCAM1*, is not detected in our CSF-CTCs (Aceto et al., 2014; Szczerba et al., 2019). CSF-CTC cell adhesion characteristics could enhance the CSF tumor cluster formation and the migration of CTCs, which is a potential target for tumor therapy strategies.

Heterogeneity of LUAD-LM tumor cells in CSF samples.

Spatial and temporal heterogeneity. The LM locations in the brain are quite diverse in all patients examined. Adaptation to different local brain environment may contribute to the gene expression heterogeneity of CSF-CTCs among patients. The scRNA-seq approach could also detect potential temporal heterogeneity during tumor progression. We obtained CSF samples from two different time points for both patient P1 (P1-1 vs. P1-2) and P8 (P8-1 vs. P8-2). Collected within a 2-month time interval, P1-1 and P1-2 have similar single-cell expression profiles and CSF-CTCs from both datasets form a single homogeneous cluster (Figure 2A), indicating similar transcriptome patterns. In contrast, P8-1 and P8-2 were collected 6-month time apart as the patient's condition worsened. Although they are still in the same cluster, we observed clear separation of single-cell expression profiles on the t-SNE plot, which may correspond to the transcriptome signatures of LM progression (Figure 5A). Systematic sampling over a time course in future studies will allow better characterization of the spatial and temporal heterogeneity.

Gene expression basis of CTC heterogeneity. One major advantage of single-cell RNA-seq approach is the ability to characterize the expression variation among individual cells. We found significant greater expression heterogeneity among patient CSF samples than among cells within patient samples, suggesting the need of personalized diagnosis and expression profiling. Differential gene expression analysis revealed the following gene families vary across patient CSF-CTC samples (Figure 3D). Serpins are highly expressed in lung cancer brain metastatic cells, promoting tumor cell survival and metastases through the inhibition of plasmin generation (Valiente et al., 2014). Galectins mediate cell-cell and cell-ECM interactions and modulate cell signaling and the tumor microenvironment, which could influence the tumor invasion, migration,

metastasis and progression (Chang et al., 2017; Compagno et al., 2014; Elola et al., 2007). Claudins (CLDN) are a family of integral membrane proteins located in epithelial cells in tight junctions bearing tissues. Their expression was found in many lung cancer histological types, and overexpression could results in tumor spreading (Soini, 2012). In addition, S100 family were highly expressed in P6, whereas P7 has elevated expression of many ribosomal protein genes (RPL and RPS). These different genes and gene families all contribute to the heterogeneity of LUAD-LM tumor cells.

Mutational profile heterogeneity. In addition to tumor location in leptomeningeal, the mutational profiles of tumor cells may also contribute to the heterogeneity. Mutations in cerebrospinal fluid cell-free DNA (CSF cfDNA) were detected by next-generation sequencing (Table S4). The discovery of common driver mutations and the development of targeted therapies dramatically improved the intracranial efficacy resulting in prolonged survival. Activating mutations of the epidermal growth factor receptor (EGFR) and anaplastic lymphoma kinase (ALK) rearrangements are key in the development of brain metastases (Li et al., 2018; Zheng et al., 2019). P1 belongs to this category with EGFR (19del) and ALK (Arg1192Trp) mutation, along with TP53 (Trp53Ter) mutation and low frequency of KRAS mutation (Table S4). EGFR (Leu858Arg) and TP53 (Arg248Gln) mutations were found in P4, and P7 only has TP53 (Asp281Tyr) mutation. We have not detected any major CSF mutations in P2 or P5. These different mutational profiles, together with different treatment strategy, also contribute to the greater among patient heterogeneity.

Potential epigenetic heterogeneity. *SCGB3A1*, a tumor suppressor in many cancer types, loses expression due to promoter methylation (Krop et al., 2005). Interestingly, *SCGB3A1* was

only highly expressed in CSF-CTCs of P2, P6, P8, but not in other LUAD-LM patients, indicating potential methylation heterogeneity among cells and patients.

CSF immune function and cancer.

Normal CSF environment has its own specialized immune cell composition, including T cells and Monocytes but not B cells (Engelhardt and Ransohoff, 2012; Larochelle et al., 2011; Wilson et al., 2010). Our scRNA-seq results of control patient samples confirmed that composition at single-cell transcriptome level. When tumor cells infiltrate CSF, the central immune system is activated to combat the tumor cells. Efficacy of cancer immunotherapies is partly depending on the amount and properties of tumor infiltrating lymphocytes (Huang et al., 2017). A recent single-cell RNA study showed great heterogeneity within the tumor regulatory T cells (Tregs) and tumor-infiltrating CD8+ T cells, and a low ratio of "pre-exhausted" to exhausted T cells was related to worse prognosis of lung adenocarcinoma (Guo et al., 2018). Therefore, analyzing the immune cell composition in pathological CSF samples would be informative for understanding the interactions between CSF microenvironment and metastatic cancer cells. The tumor infiltrating immune cells in CSF samples is worth future research.

The metastatic process – partial EMT and stemness

EMT is a process relevant tumor invasion and metastases. It has been reported that NSCLC blood-CTCs have a dual epithelial-mesenchymal phenotype (Lecharpentier et al., 2011). Similarly, we discovered abundant expression of epithelial genes in LUAD-LM CSF-CTCs in our research. In addition, a small subset of CSF-CTCs expressed mesenchymal genes (*VIM* and *FN1*) but without the expression of EMT transcription factors which are necessary for EMT. Notably, we also identified the unexpected abundant expression of extracellular matrix (ECM)

genes in CSF-CTCs (Ting et al., 2014). Tumor stroma-derived ECM signaling plays an important role in targeting cancer cell metastasis (Zhang et al., 2013). The cell-autonomous expression of ECM genes in CSF-CTCs may contribute to the dissemination of cancer.

Our research revealed a couple major differences with NSCLC blood-CTCs. Cells with high expression of mesenchymal genes and low expression of epithelial genes are extremely rare in LUAD-LM CSF-CTCs (only 2 cells in Figure S13A). We did not observe any correlation between CD44 expression and enrichment for the mesenchymal genes (*VIM* or *FN1*) within single CSF-CTCs, suggesting that stem cell and EMT markers are not intrinsically linked in CSF-CTCs (Figure S13B). The advancement of the CSF-CTC expression files and comparison with NSCLC blood-CTCs will provide a much better understanding of the mechanisms of LUAD-LM.

The power of single-cell RNA sequencing in CSF samples for the diagnosis of CUP origin.

CUP is a well-recognized clinical disorder, which accounts for 3-5% of all malignant epithelial tumors with poor prognosis due to treatment of a non-selective empirical therapy. Identification of the primary tumor type will greatly inform the treatment strategies, but it is extremely challenging. It is estimated that about 80% CUP cases are metastatic adenocarcinoma (Pavlidis and Pentheroudakis, 2012). Our studies also focused on this most common CUP, by enrolling one CUP patient of metastatic adenocarcinoma (P8).

In order to pinpoint the organ and tissue of origin, patient history, physical examination, serum markers, histological data and state-of-the-art imaging results have been examined and the primary origin remains inconclusive. For CUP patients with leptomeningeal metastases only, the CSF-CTCs are the available tissue sample for the diagnosis of the primary origin of CUP. Since

P8 is a CUP case with multi-site metastases, we have the biopsy of left lymph nodes to perform immunohistochemistry (IHC) to pair with our single-cell RNA-seq data. The scRNA-seq data in CSF-CTCs and IHC results in lymph nodes revealed the epithelial origin and it is less likely to have lung, prostate, gastrointestinal or liver origin, providing crucial diagnostic information for this patient. The cluster-defined genes, *PIP* and *ANKRD30A*, are exclusively expressed in P8 compared to other LUAD-LM CSF samples, indicating sufficient evidence to diagnose the primary site as breast cancer, sweat/salivary gland cancer and prostate cancer, though a definitive conclusion could not be made because the patient passed away and refused autopsy.

The scRNA-seq results provided promising directions for diagnosis. Further investigation has been made on the possibilities of prostate and salivary gland cancer, and no evidence has been found despite extensive imaging examinations. Interestingly, when we checked the expression of *SCGB2A2*, a classical marker of breast cancer, 7 CTCs from P8 have high expression levels (Figure 5C) whereas other LUAD-LM patients have little to no expression in CSF-CTCs. Currently, breast cancer origin is the best suggestive diagnosis we could make based all evidence we obtained.

Immunohistochemical markers are the most important diagnostic tools in establishing tissue origin of CUP, but scRNA-seq has far better sensitivity, especially for CSF samples with low tumor burden. The successful detection of 7 cells expressing SCGB2A2 in patient P8 is one example of the advantage of single-cell sequencing over bulk RNA-seq or IHC. As the first CUP case with scRNA-seq data in CSF-CTCs, we were able to achieve a comprehensive characterization of the transcriptome pattern in every P8 tumor cell, as well as the discovery of potential biomarker expressed at a low frequency in specific cells. With continuous advancement of scRNA-seq technology and deceasing sequencing cost, additional scRNA-seq datasets will be

available for breast cancer, sweat and salivary gland cancer, and prostate cancer. In the near future, we will be able to build transcriptome databases of multiple CUP cases and provide speedy and accurate diagnosis for CUP cases to benefit the cancer patients.

The transcriptome signature of CSF-CTCs in the CUP case.

Traditional diagnostic methods rely on CSF cytology and the examination of known marker genes. Our scRNA-seq approach can only profile the classic markers with great sensitivity and precision, but also discover novel markers which are not known to be associated with the disease. The CSF-CTCs of P8 showed low-proliferative and high-epithelial signatures as LUAD-LM patients. Unlike LUAD-LM CTCs, we found that CSF-CTCs were enriched for the stem-cell-associated gene *PROM1* and did not upregulate ECM receptor interaction pathway.

We also investigated other important P8 cluster-defined genes. Among them *DCD*, *MSMB*, *ZG16B* and *TFF1* have been previously reported to associate with many types of metastatic adenocarcinoma (Figure 5B). *DCD* (dermcidin), a constitutively expressed gene in eccrine sweat glands (Burian and Schittek, 2015; Schittek, 2012), is expressed in 52 of 442 P8 CSF-CTCs. It is reported that peptides processed from the dermcidin precursor exhibit a range of other biological functions in neuronal and cancer cells (Schittek, 2012; Stewart et al., 2008). *MSMB* (microseminoprotein beta, 93/442 in P8 CTCs) is one of the most abundant proteins in semen and is correlated with prostate cancer progression (Emami et al., 2019; Sutcliffe et al., 2014; Whitaker et al., 2010). *MSMB* was enrolled in “STHLM3 model” for diagnosis of prostate cancer (Gronberg et al., 2015) and SNP rs10993994 was associated with prostate cancer risk (Eeles et al., 2008; Lou et al., 2009; Shui et al., 2014; Thomas et al., 2008). *ZG16B/PAUF* (pancreatic adenocarcinoma-upregulated factor, 214/442 in P8 CTCs) is a secreted protein with

crucial role in pancreatic ductal adenocarcinoma (Kim et al., 2009; Kim et al., 2013; Lee et al., 2010; Park et al., 2011). *ZG16B/PAUF* has also been shown to be present in other cancer types including epithelial ovarian cancer (Choi et al., 2018), cervical carcinoma (Kim et al., 2018), colorectal cancer (Escudero-Paniagua et al., 2019) and oral squamous cell carcinoma (Sasahira et al., 2017). The trefoil factor (TFF) family of proteins, in particular *TFF1*, are secreted by the mucus-secreting cells and plays essential roles mainly in breast and gastric cancer (Perry et al., 2008). In addition, other P8 cluster-defined genes shown in Figure S12 are valuable candidates for future research.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (Grant No. 2017YFA0103902), the National Natural Science Foundation of China (Grant No. 31771283), the Fundamental Research Funds for the Central Universities (Grant No. 22120190210), the Innovation Group Project of Shanghai Municipal Health Commission (2019CXJQ03) and an Auburn University Research Initiative in Cancer (ARUIC) Research Grant. M.G. is supported by the Program for Shanghai Municipal Leading Talent (2015). X.W. is supported by the USDA National Institute of Food and Agriculture (Hatch project 1018100) and a generous laboratory start-up fund from Auburn University College of Veterinary Medicine. Y.Z. is supported by Auburn University Presidential Graduate Research Fellowship.

References

Aceto, N., Bardia, A., Miyamoto, D. T., Donaldson, M. C., Wittner, B. S., Spencer, J. A., Yu, M., Pely, A., Engstrom, A., Zhu, H., *et al.* (2014). Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell *158*, 1110-1122.

Alberg, A. J., Brock, M. V., and Samet, J. M. (2005). Epidemiology of lung cancer: looking to the future. J Clin Oncol *23*, 3175-3185.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. In.

Bacchi, C. E., and Gown, A. M. (1991). Distribution and pattern of expression of villin, a gastrointestinal-associated cytoskeletal protein, in human carcinomas: a study employing paraffin-embedded tissue. Lab Invest *64*, 418-424.

Barnholtz-Sloan, J. S., Sloan, A. E., Davis, F. G., Vigneau, F. D., Lai, P., and Sawaya, R. E. (2004). Incidence proportions of brain metastases in patients diagnosed (1973 to 2001) in the Metropolitan Detroit Cancer Surveillance System. J Clin Oncol *22*, 2865-2872.

Beauchemin, N., and Arabzadeh, A. (2013). Carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) in cancer progression and metastasis. Cancer Metastasis Rev *32*, 643-671.

Bertolini, G., Roz, L., Perego, P., Tortoreto, M., Fontanella, E., Gatti, L., Pratesi, G., Fabbri, A., Andriani, F., Tinelli, S., *et al.* (2009). Highly tumorigenic lung cancer CD133+ cells display stem-like features and are spared by cisplatin treatment. Proc Natl Acad Sci U S A *106*, 16281-16286.

Bleau, A. M., Huse, J. T., and Holland, E. C. (2009). The ABCG2 resistance network of glioblastoma. Cell Cycle *8*, 2936-2944.

Bohinski, R. J., Di Lauro, R., and Whitsett, J. A. (1994). The lung-specific surfactant protein B gene promoter is a target for thyroid transcription factor 1 and hepatocyte nuclear factor 3, indicating common factors for organ-specific gene expression along the foregut axis. Mol Cell Biol *14*, 5671-5681.

Boire, A., Zou, Y., Shieh, J., Macalinao, D. G., Pentsova, E., and Massague, J. (2017). Complement Component 3 Adapts the Cerebrospinal Fluid for Leptomeningeal Metastasis. Cell *168*, 1101-1113 e1113.

Bruno, M. D., Bohinski, R. J., Huelsman, K. M., Whitsett, J. A., and Korfhagen, T. R. (1995). Lung cell-specific expression of the murine surfactant protein A (SP-A) gene is mediated by interactions between the SP-A promoter and thyroid transcription factor-1. J Biol Chem *270*, 6531-6536.

Burian, M., and Schittek, B. (2015). The secrets of dermcidin action. Int J Med Microbiol *305*, 283-286.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology *36*, 411-+.

Chang, W. A., Tsai, M. J., Kuo, P. L., and Hung, J. Y. (2017). Role of galectins in lung cancer. Oncol Lett *14*, 5077-5084.

Chen, E. I., Hewel, J., Krueger, J. S., Tiraby, C., Weber, M. R., Kralli, A., Becker, K., Yates, J. R., 3rd, and Felding-Habermann, B. (2007). Adaptation of energy metabolism in breast cancer brain metastases. Cancer Res *67*, 1472-1486.

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X. Q., and He, J. (2016). Cancer statistics in China, 2015. CA Cancer J Clin *66*, 115-132.

Cheng, H., and Perez-Soler, R. (2018). Leptomeningeal metastases in non-small-cell lung cancer. Lancet Oncol *19*, e43-e55.

Chiba, Y., Kurotani, R., Kusakabe, T., Miura, T., Link, B. W., Misawa, M., and Kimura, S. (2006). Uteroglobin-related protein 1 expression suppresses allergic airway inflammation in mice. Am J Respir Crit Care Med *173*, 958-964.

Choi, C. H., Kang, T. H., Song, J. S., Kim, Y. S., Chung, E. J., Ylaya, K., Kim, S., Koh, S. S., Chung, J. Y., Kim, J. H., and Hewitt, S. M. (2018). Elevated expression of pancreatic adenocarcinoma upregulated factor (PAUF) is associated with poor prognosis and chemoresistance in epithelial ovarian cancer. Sci Rep *8*, 12161.

Compagno, D., Jaworski, F. M., Gentilini, L., Contrufo, G., Gonzalez Perez, I., Elola, M. T., Pregi, N., Rabinovich, G. A., and Laderach, D. J. (2014). Galectins: major signaling modulators inside and outside the cell. Curr Mol Med *14*, 630-651.

Craig, S. E., and Brady-Kalnay, S. M. (2011). Cancer cells cut homophilic cell adhesion molecules and run. Cancer Res *71*, 303-309.

De Mattos-Arruda, L., Mayor, R., Ng, C. K. Y., Weigelt, B., Martinez-Ricarte, F., Torrejon, D., Oliveira, M., Arias, A., Raventos, C., Tang, J., *et al.* (2015). Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. Nat Commun *6*, 8839.

Di Loreto, C., Di Lauro, V., Puglisi, F., Damante, G., Fabbro, D., and Beltrami, C. A. (1997). Immunocytochemical expression of tissue specific transcription factor-1 in lung carcinoma. J Clin Pathol *50*, 30-32.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Eeles, R. A., Kote-Jarai, Z., Giles, G. G., Olama, A. A., Guy, M., Jugurnauth, S. K., Mulholland, S., Leongamornlert, D. A., Edwards, S. M., Morrison, J., *et al.* (2008). Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet *40*, 316-321.

Elola, M. T., Wolfenstein-Todel, C., Troncoso, M. F., Vasta, G. R., and Rabinovich, G. A. (2007). Galectins: matricellular glycan-binding proteins linking cell adhesion, migration, and survival. Cell Mol Life Sci *64*, 1679-1700.

Emami, N. C., Kachuri, L., Meyers, T. J., Das, R., Hoffman, J. D., Hoffmann, T. J., Hu, D., Shan, J., Feng, F. Y., Ziv, E., *et al.* (2019). Association of imputed prostate cancer transcriptome with disease risk reveals novel mechanisms. Nat Commun *10*, 3107.

Engelhardt, B., and Ransohoff, R. M. (2012). Capture, crawl, cross: the T cell code to breach the blood-brain barriers. Trends Immunol *33*, 579-589.

Escudero-Paniagua, B., Bartolome, R. A., Rodriguez, S., de Los Rios, V., Pintado, L., Jaen, M., Lafarga, M., Fernandez-Acenero, M. J., and Casal, J. I. (2019). PAUF/ZG16B promotes colorectal cancer progression through alterations of the mitotic functions and the Wnt/beta-catenin pathway. Carcinogenesis.

Ewels, P., Magnusson, M., Lundin, S., and Kaller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047-3048.

Farhadian, S. F., Mehta, S. S., Zografou, C., Robertson, K., Price, R. W., Pappalardo, J., Chiarella, J., Hafler, D. A., and Spudich, S. S. (2018). Single-cell RNA sequencing reveals microglia-like cells in cerebrospinal fluid during virologically suppressed HIV. JCI Insight *3*.

Filmus, J., and Capurro, M. (2013). Glypican-3: a marker and a therapeutic target in hepatocellular carcinoma. FEBS J *280*, 2471-2476.

Forner, D., Bullock, M., Manders, D., Wallace, T., Chin, C. J., Johnson, L. B., Rigby, M. H., Trites, J. R., Taylor, M. S., and Hart, R. D. (2018). Secretory carcinoma: the eastern Canadian experience and literature review. J Otolaryngol Head Neck Surg *47*, 69.

Giancotti, F. G., and Ruoslahti, E. (1999). Integrin signaling. Science *285*, 1028-1032.

Giladi, A., and Amit, I. (2017). Immunology, one cell at a time. Nature *547*, 27-29.

Glasser, S. W., Eszterhas, S. K., Detmer, E. A., Maxfield, M. D., and Korfhagen, T. R. (2005). The murine SP-C promoter directs type II cell-specific expression in transgenic mice. Am J Physiol Lung Cell Mol Physiol *288*, L625-632.

Gleissner, B., and Chamberlain, M. C. (2006). Neoplastic meningitis. Lancet Neurol *5*, 443-452.

Green, B. J., Saberi Safaei, T., Mepham, A., Labib, M., Mohamadi, R. M., and Kelley, S. O. (2016). Beyond the Capture of Circulating Tumor Cells: Next-Generation Devices and Materials. Angew Chem Int Ed Engl *55*, 1252-1265.

Grinberg-Rashi, H., Ofek, E., Perelman, M., Skarda, J., Yaron, P., Hajduch, M., Jacob-Hirsch, J., Amariglio, N., Krupsky, M., Simansky, D. A., *et al.* (2009). The expression of three genes in primary non-small cell lung cancer is associated with metastatic spread to the brain. Clin Cancer Res *15*, 1755-1761.

Gronberg, H., Adolfsson, J., Aly, M., Nordstrom, T., Wiklund, P., Brandberg, Y., Thompson, J., Wiklund, F., Lindberg, J., Clements, M., *et al.* (2015). Prostate cancer screening in men aged 50-69 years (STHLM3): a prospective population-based diagnostic study. Lancet Oncol *16*, 1667-1676.

Grossman, S. A., and Krabak, M. J. (1999). Leptomeningeal carcinomatosis. Cancer Treat Rev *25*, 103-119.

Guo, X., Zhang, Y., Zheng, L., Zheng, C., Song, J., Zhang, Q., Kang, B., Liu, Z., Jin, L., Xing, R., *et al.* (2018). Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med *24*, 978-985.

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 7.

Herbst, R. S., Heymach, J. V., and Lippman, S. M. (2008). Lung cancer. N Engl J Med *359*, 1367-1380.

Höftberger, R., Mader, S. A., and Reindl, M. (2015). Cerebrospinal Fluid in Clinical Neurology. In, pp. 143-158.

Huang, A. C., Postow, M. A., Orlowski, R. J., Mick, R., Bengsch, B., Manne, S., Xu, W., Harmon, S., Giles, J. R., Wenz, B., *et al.* (2017). T-cell invigoration to tumour burden ratio associated with anti-PD-1 response. Nature *545*, 60-65.

Jager, D., Karbach, J., Pauligk, C., Seil, I., Frei, C., Chen, Y. T., Old, L. J., Knuth, A., and Jager, E. (2005). Humoral and cellular immune responses against the breast cancer antigen NY-BR-1: definition of two HLA-A2 restricted peptide epitopes. Cancer Immun *5*, 11.

Jiang, B. Y., Li, Y. S., Guo, W. B., Zhang, X. C., Chen, Z. H., Su, J., Zhong, W. Z., Yang, X. N., Yang, J. J., Shao, Y., *et al.* (2017). Detection of Driver and Resistance Mutations in Leptomeningeal Metastases of NSCLC by Next-Generation Sequencing of Cerebrospinal Fluid Circulating Tumor Cells. Clin Cancer Res *23*, 5480-5488.

Julian, T. B., Blumencranz, P., Deck, K., Whitworth, P., Berry, D. A., Berry, S. M., Rosenberg, A., Chagpar, A. B., Reintgen, D., Beitsch, P., *et al.* (2008). Novel intraoperative molecular test for sentinel lymph node metastases in patients with early-stage breast cancer. J Clin Oncol *26*, 3338-3345.

Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Sci *28*, 1947-1951.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research *28*, 27-30.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. Nucleic Acids Research *47*, D590-D595.

Kelly, S. E., Bachurski, C. J., Burhans, M. S., and Glasser, S. W. (1996). Transcription of the lung-specific surfactant protein C gene is mediated by thyroid transcription factor 1. J Biol Chem *271*, 6881-6888.

Kim, J., Chung, J. Y., Kim, T. J., Lee, J. W., Kim, B. G., Bae, D. S., Choi, C. H., and Hewitt, S. M. (2018). Genomic Network-Based Analysis Reveals Pancreatic Adenocarcinoma Up-Regulating Factor-Related Prognostic Markers in Cervical Carcinoma. Front Oncol *8*, 465.

Kim, K. T., Lee, H. W., Lee, H. O., Kim, S. C., Seo, Y. J., Chung, W., Eum, H. H., Nam, D. H., Kim, J., Joo, K. M., and Park, W. Y. (2015). Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol *16*, 127.

Kim, M. Y., Go, H., Koh, J., Lee, K., Min, H. S., Kim, M. A., Jeon, Y. K., Lee, H. S., Moon, K. C., Park, S. Y., *et al.* (2014). Napsin A is a useful marker for metastatic adenocarcinomas of pulmonary origin. Histopathology *65*, 195-206.

Kim, S. A., Lee, Y., Jung, D. E., Park, K. H., Park, J. Y., Gang, J., Jeon, S. B., Park, E. C., Kim, Y. G., Lee, B., *et al.* (2009). Pancreatic adenocarcinoma up-regulated factor (PAUF), a novel up-regulated secretory protein in pancreatic ductal adenocarcinoma. Cancer Sci *100*, 828-836.

Kim, S. J., Lee, Y., Kim, N. Y., Hwang, Y., Hwang, B., Min, J. K., and Koh, S. S. (2013). Pancreatic adenocarcinoma upregulated factor, a novel endothelial activator, promotes angiogenesis and vascular permeability. Oncogene *32*, 3638-3647.

Krop, I., Parker, M. T., Bloushtain-Qimron, N., Porter, D., Gelman, R., Sasaki, H., Maurer, M., Terry, M. B., Parsons, R., and Polyak, K. (2005). HIN-1, an inhibitor of cell growth, invasion, and AKT activation. Cancer Res *65*, 9659-9669.

Kruger, F. (2012). TrimGalore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. In.

Kurotani, R., Kumaki, N., Naizhen, X., Ward, J. M., Linnoila, R. I., and Kimura, S. (2011). Secretoglobin 3A2/uteroglobin-related protein 1 is a novel marker for pulmonary carcinoma in mice and humans. Lung Cancer *71*, 42-48.

Kurotani, R., Tomita, T., Yang, Q., Carlson, B. A., Chen, C., and Kimura, S. (2008). Role of secretoglobin 3A2 in lung development. Am J Respir Crit Care Med *178*, 389-398.

Lacroix, M. (2006). Significance, detection and markers of disseminated breast cancer cells. Endocr Relat Cancer *13*, 1033-1067.

Lambert, A. W., Pattabiraman, D. R., and Weinberg, R. A. (2017). Emerging Biological Principles of Metastasis. Cell *168*, 670-691.

Larochelle, C., Alvarez, J. I., and Prat, A. (2011). How do immune cells overcome the blood-brain barrier in multiple sclerosis? FEBS Lett *585*, 3770-3780.

Lassman, A. B., and DeAngelis, L. M. (2003). Brain metastases. Neurol Clin *21*, 1-23, vii.

Lecharpentier, A., Vielh, P., Perez-Moreno, P., Planchard, D., Soria, J. C., and Farace, F. (2011). Detection of circulating tumour cells with a hybrid (epithelial/mesenchymal) phenotype in patients with metastatic non-small cell lung cancer. Br J Cancer *105*, 1338-1341.

Lee, Y., Kim, S. J., Park, H. D., Park, E. H., Huang, S. M., Jeon, S. B., Kim, J. M., Lim, D. S., and Koh, S. S. (2010). PAUF functions in the metastasis of human pancreatic cancer cells and upregulates CXCR4 expression. Oncogene *29*, 56-67.

Leung, E. L., Fiscus, R. R., Tung, J. W., Tin, V. P., Cheng, L. C., Sihoe, A. D., Fink, L. M., Ma, Y., and Wong, M. P. (2010). Non-small cell lung cancer cells expressing CD44 are enriched for stem cell-like properties. PLoS One *5*, e14062.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Proc, G. P. D. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, Y. S., Jiang, B. Y., Yang, J. J., Zhang, X. C., Zhang, Z., Ye, J. Y., Zhong, W. Z., Tu, H. Y., Chen, H. J., Wang, Z., *et al.* (2018). Unique genetic profiles from cerebrospinal fluid cell-free DNA in leptomeningeal metastases of EGFR-mutant non-small-cell lung cancer: a new medium of liquid biopsy. Ann Oncol *29*, 945-952.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

Lin, X., Fleisher, M., Rosenblum, M., Lin, O., Boire, A., Briggs, S., Bensman, Y., Hurtado, B., Shagabayeva, L., DeAngelis, L. M., *et al.* (2017). Cerebrospinal fluid circulating tumor cells: a novel tool to diagnose leptomeningeal metastases from epithelial tumors. Neuro Oncol *19*, 1248-1254.

Linnoila, R. I., Szabo, E., DeMayo, F., Witschi, H., Sabourin, C., and Malkinson, A. (2000). The role of CC10 in pulmonary carcinogenesis: from a marker to tumor suppression. Ann N Y Acad Sci *923*, 249-267.

Liu, X., Taftaf, R., Kawaguchi, M., Chang, Y. F., Chen, W., Entenberg, D., Zhang, Y., Gerratana, L., Huang, S., Patel, D. B., *et al.* (2019). Homophilic CD44 Interactions Mediate Tumor Cell Aggregation and Polyclonal Metastasis in Patient-Derived Breast Cancer Models. Cancer Discov *9*, 96-113.

Lou, H., Yeager, M., Li, H., Bosquet, J. G., Hayes, R. B., Orr, N., Yu, K., Hutchinson, A., Jacobs, K. B., Kraft, P., *et al.* (2009). Fine mapping and functional analysis of a common variant in MSMB on chromosome 10q11.2 associated with prostate cancer susceptibility. Proc Natl Acad Sci U S A *106*, 7933-7938.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*.

Lowery, F. J., and Yu, D. (2017). Brain metastasis: Unique challenges and open opportunities. Biochim Biophys Acta Rev Cancer *1867*, 49-57.

Ma, K. Y., Schonnesen, A. A., Brock, A., Van Den Berg, C., Eckhardt, S. G., Liu, Z., and Jiang, N. (2019). Single-cell RNA sequencing of lung adenocarcinoma reveals heterogeneity of immune response-related genes. JCI Insight *4*.

MacDonagh, L., Gray, S. G., Breen, E., Cuffe, S., Finn, S. P., O'Byrne, K. J., and Barr, M. P. (2016). Lung cancer stem cells: The root of resistance. Cancer Lett *372*, 147-156.

Mao, S., Zhang, Q., Li, H., Zhang, W., Huang, Q., Khan, M., and Lin, J. M. (2018). Adhesion analysis of single circulating tumor cells on a base layer of endothelial cells using open microfluidics. Chem Sci *9*, 7694-7699.

Markou, A., Strati, A., Malamos, N., Georgoulias, V., and Lianidou, E. S. (2011). Molecular characterization of circulating tumor cells in breast cancer by a liquid bead array hybridization assay. Clin Chem *57*, 421-430.

Mengoli, M. C., Longo, F. R., Fraggetta, F., Cavazza, A., Dubini, A., Ali, G., Guddo, F., Gilioli, E., Bogina, G., Nannini, N., *et al.* (2018). The 2015 World Health Organization Classification of lung tumors: new entities since the 2004 Classification. Pathologica *110*, 39-67.

Mergenthaler, P., Lindauer, U., Dienel, G. A., and Meisel, A. (2013). Sugar for the brain: the role of glucose in physiological and pathological brain function. Trends Neurosci *36*, 587-597.

Miyamoto, D. T., Ting, D. T., Toner, M., Maheswaran, S., and Haber, D. A. (2016). Single-Cell Analysis of Circulating Tumor Cells as a Window into Tumor Heterogeneity. Cold Spring Harb Symp Quant Biol *81*, 269-274.

Miyamoto, D. T., Zheng, Y., Wittner, B. S., Lee, R. J., Zhu, H., Broderick, K. T., Desai, R., Fox, D. B., Brannigan, B. W., Trautwein, J., *et al.* (2015). RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science *349*, 1351-1356.

Nakamura, N., Miyagi, E., Murata, S., Kawaoi, A., and Katoh, R. (2002). Expression of thyroid transcription factor-1 in normal and neoplastic lung tissues. Mod Pathol *15*, 1058-1067.

Nayak, L., Fleisher, M., Gonzalez-Espinoza, R., Lin, O., Panageas, K., Reiner, A., Liu, C. M., Deangelis, L. M., and Omuro, A. (2013). Rare cell capture technology for the diagnosis of leptomeningeal metastasis in solid tumors. Neurology *80*, 1598-1605; discussion 1603.

Paliouras, M., Borgono, C., and Diamandis, E. P. (2007). Human tissue kallikreins: the cancer biomarker family. Cancer Lett *249*, 61-79.

Park, H. D., Lee, Y., Oh, Y. K., Jung, J. G., Park, Y. W., Myung, K., Kim, K. H., Koh, S. S., and Lim, D. S. (2011). Pancreatic adenocarcinoma upregulated factor promotes metastasis by regulating TLR/CXCR4 activation. Oncogene *30*, 201-211.

Patel, M., Lu, L., Zander, D. S., Sreerama, L., Coco, D., and Moreb, J. S. (2008). ALDH1A1 and ALDH3A1 expression in lung cancers: correlation with histologic type and potential precursors. Lung Cancer *59*, 340-349.

Pavlidis, N., and Pentheroudakis, G. (2012). Cancer of unknown primary site. Lancet *379*, 1428-1435.

Peng, J., Sun, B. F., Chen, C. Y., Zhou, J. Y., Chen, Y. S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G. S., *et al.* (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res *29*, 725-738.

Perry, J. K., Kannan, N., Grandison, P. M., Mitchell, M. D., and Lobie, P. E. (2008). Are trefoil factors oncogenic? Trends Endocrinol Metab *19*, 74-81.

Peters, S., Bexelius, C., Munk, V., and Leighl, N. (2016). The impact of brain metastasis on quality of life, resource utilization and survival in patients with non-small-cell lung cancer. Cancer Treat Rev *45*, 139-162.

Picelli, S., Bjorklund, A. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods *10*, 1096-1098.

Poornima, S., Christian, C., and Kresch, M. J. (2002). Developmental regulation of SP-A receptor in fetal rat lung. Lung *180*, 33-46.

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C. L., Mroz, E. A., Emerick, K. S., *et al.* (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell *171*, 1611-1624 e1624.

Ray, M. K., Chen, C. Y., Schwartz, R. J., and DeMayo, F. J. (1996). Transcriptional regulation of a mouse Clara cell-specific protein (mCC10) gene by the NKx transcription factor family members thyroid transciption factor 1 and cardiac muscle-specific homeobox protein (CSX). Mol Cell Biol *16*, 2056-2064.

Razavi, P., Li, B. T., Brown, D. N., Jung, B., Hubbell, E., Shen, R., Abida, W., Juluru, K., De Bruijn, I., Hou, C., *et al.* (2019). High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. Nat Med *25*, 1928-1937.

Reinholz, M. M., Kitzmann, K. A., Tenner, K., Hillman, D., Dueck, A. C., Hobday, T. J., Northfelt, D. W., Moreno-Aspitia, A., Roy, V., LaPlant, B., *et al.* (2011). Cytokeratin-19 and mammaglobin gene expression in circulating tumor cells from metastatic breast cancer patients enrolled in North Central Cancer Treatment Group trials, N0234/336/436/437. Clin Cancer Res *17*, 7183-7193.

Remon, J., Le Rhun, E., and Besse, B. (2017). Leptomeningeal carcinomatosis in non-small cell lung cancer patients: A continuing challenge in the personalized treatment era. Cancer Treat Rev *53*, 128-137.

Saitoh, M. (2018). Involvement of partial EMT in cancer progression. J Biochem *164*, 257-264.

Salmaninejad, A., Zamani, M. R., Pourvahedi, M., Golchehre, Z., Hosseini Bereshneh, A., and Rezaei, N. (2016). Cancer/Testis Antigens: Expression, Regulation, Tumor Invasion, and Use in Immunotherapy of Cancers. Immunol Invest *45*, 619-640.

Sankey, E. W., Tsvankin, V., Grabowski, M. M., Nayar, G., Batich, K. A., Risman, A., Champion, C. D., Salama, A. K. S., Goodwin, C. R., and Fecci, P. E. (2019). Operative and peri-operative considerations in the management of brain metastasis. Cancer Med *8*, 6809-6831.

Sasahira, T., Kurihara, M., Nishiguchi, Y., Nakashima, C., Kirita, T., and Kuniyasu, H. (2017). Pancreatic adenocarcinoma up-regulated factor has oncogenic functions in oral squamous cell carcinoma. Histopathology *70*, 539-548.

Sasaki, S., Yoshioka, Y., Ko, R., Katsura, Y., Namba, Y., Shukuya, T., Kido, K., Iwakami, S., Tominaga, S., and Takahashi, K. (2016). Diagnostic significance of cerebrospinal fluid EGFR mutation analysis for leptomeningeal metastasis in non-small-cell lung cancer patients harboring an active EGFR mutation following gefitinib therapy failure. Respir Investig *54*, 14-19.

Schittek, B. (2012). The multiple facets of dermcidin in cell survival and host defense. J Innate Immun *4*, 349-360.

Schouten, L. J., Rutten, J., Huveneers, H. A., and Twijnstra, A. (2002). Incidence of brain metastases in a cohort of patients with carcinoma of the breast, colon, kidney, and lung and melanoma. Cancer *94*, 2698-2705.

Seil, I., Frei, C., Sultmann, H., Knauer, S. K., Engels, K., Jager, E., Zatloukal, K., Pfreundschuh, M., Knuth, A., Tseng-Chen, Y., *et al.* (2007). The differentiation antigen NY-BR-1 is a potential target for antibody-based therapies in breast cancer. Int J Cancer *120*, 2635-2642.

Sevenich, L., and Joyce, J. A. (2014). Pericellular proteolysis in cancer. Genes Dev *28*, 2331-2347.

Shao, C., Sullivan, J. P., Girard, L., Augustyn, A., Yenerall, P., Rodriguez-Canales, J., Liu, H., Behrens, C., Shay, J. W., Wistuba, II, and Minna, J. D. (2014). Essential role of aldehyde dehydrogenase 1A3 for the maintenance of non-small cell lung cancer stem cells is associated with the STAT3 pathway. Clin Cancer Res *20*, 4154-4166.

Shui, I. M., Lindstrom, S., Kibel, A. S., Berndt, S. I., Campa, D., Gerke, T., Penney, K. L., Albanes, D., Berg, C., Bueno-de-Mesquita, H. B., *et al.* (2014). Prostate cancer (PCa) risk variants and risk of fatal PCa in the National Cancer Institute Breast and Prostate Cancer Cohort Consortium. Eur Urol *65*, 1069-1075.

Sick, E., Jeanne, A., Schneider, C., Dedieu, S., Takeda, K., and Martiny, L. (2012). CD47 update: a multifaceted actor in the tumour microenvironment of potential therapeutic interest. Br J Pharmacol *167*, 1415-1430.

Siegel, R. L., Miller, K. D., and Jemal, A. (2016). Cancer statistics, 2016. CA Cancer J Clin *66*, 7-30.

Sin, D. D., Tammemagi, C. M., Lam, S., Barnett, M. J., Duan, X., Tam, A., Auman, H., Feng, Z., Goodman, G. E., Hanash, S., and Taguchi, A. (2013). Pro-surfactant protein B as a biomarker for lung cancer prediction. J Clin Oncol *31*, 4536-4543.

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Research *27*, 491-499.

Soini, Y. (2012). Tight junctions in lung cancer and lung metastasis: a review. Int J Clin Exp Pathol *5*, 126-136.

Stewart, G. D., Skipworth, R. J., Ross, J. A., Fearon, K., and Baracos, V. E. (2008). The dermcidin gene in cancer: role in cachexia, carcinogenesis and tumour cell survival. Curr Opin Clin Nutr Metab Care *11*, 208-213.

Strayer, M., Savani, R. C., Gonzales, L. W., Zaman, A., Cui, Z., Veszelovszky, E., Wood, E., Ho, Y. S., and Ballard, P. L. (2002). Human surfactant protein B promoter in transgenic mice: temporal, spatial, and stimulus-responsive regulation. Am J Physiol Lung Cell Mol Physiol *282*, L394-404.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y. H., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-+.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA *102*, 15545-15550.

Sullivan, J. P., Spinola, M., Dodge, M., Raso, M. G., Behrens, C., Gao, B., Schuster, K., Shao, C., Larsen, J. E., Sullivan, L. A., *et al.* (2010). Aldehyde dehydrogenase activity selects for lung adenocarcinoma stem cells dependent on notch signaling. Cancer Res *70*, 9937-9948.

Sutcliffe, S., De Marzo, A. M., Sfanos, K. S., and Laurence, M. (2014). MSMB variation and prostate cancer risk: clues towards a possible fungal etiology. Prostate *74*, 569-578.

Szczerba, B. M., Castro-Giner, F., Vetter, M., Krol, I., Gkountela, S., Landin, J., Scheidmann, M. C., Donato, C., Scherrer, R., Singer, J., *et al.* (2019). Neutrophils escort circulating tumour cells to enable cell cycle progression. Nature *566*, 553-557.

Tafreshi, N. K., Enkemann, S. A., Bui, M. M., Lloyd, M. C., Abrahams, D., Huynh, A. S., Kim, J., Grobmyer, S. R., Carter, W. B., Vagner, J., *et al.* (2011). A mammaglobin-A targeting agent for noninvasive detection of breast cancer metastasis in lymph nodes. Cancer Res *71*, 1050-1059.

Theocharis, A. D., Skandalis, S. S., Gialeli, C., and Karamanos, N. K. (2016). Extracellular matrix structure. Adv Drug Deliv Rev *97*, 4-27.

Theurillat, J. P., Zurrer-Hardi, U., Varga, Z., Storz, M., Probst-Hensch, N. M., Seifert, B., Fehr, M. K., Fink, D., Ferrone, S., Pestalozzi, B., *et al.* (2007). NY-BR-1 protein expression in breast carcinoma: a mammary gland differentiation antigen as target for cancer immunotherapy. Cancer Immunol Immunother *56*, 1723-1731.

Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., *et al.* (2008). Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet *40*, 310-315.

Ting, D. T., Wittner, B. S., Ligorio, M., Vincent Jordan, N., Shah, A. M., Miyamoto, D. T., Aceto, N., Bersani, F., Brannigan, B. W., Xega, K., *et al.* (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep *8*, 1905-1918.

Traebert, M., Hattenhauer, O., Murer, H., Kaissling, B., and Biber, J. (1999). Expression of type II Na-P(i) cotransporter in alveolar type II cells. Am J Physiol *277*, L868-873.

Tu, Q., Wu, X., Le Rhun, E., Blonski, M., Wittwer, B., Taillandier, L., De Carvalho Bittencourt, M., and Faure, G. C. (2015). CellSearch technology applied to the detection and quantification of tumor cells in CSF of patients with lung cancer leptomeningeal metastasis. Lung Cancer *90*, 352-357.

Urbaniak, A., Jablonska, K., Podhorska-Okolow, M., Ugorski, M., and Dziegiel, P. (2018). Prolactin-induced protein (PIP)-characterization and role in breast cancer progression. Am J Cancer Res *8*, 2150-2164.

Valiente, M., Obenauf, A. C., Jin, X., Chen, Q., Zhang, X. H., Lee, D. J., Chaft, J. E., Kris, M. G., Huse, J. T., Brogi, E., and Massague, J. (2014). Serpins promote cancer cell survival and vascular co-option in brain metastasis. Cell *156*, 1002-1016.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J Mach Learn Res *9*, 2579-2605.

Varga, Z., Theurillat, J. P., Filonenko, V., Sasse, B., Odermatt, B., Jungbluth, A. A., Chen, Y. T., Old, L. J., Knuth, A., Jager, D., and Moch, H. (2006). Preferential nuclear and cytoplasmic NY-BR-1 protein expression in primary breast cancer and lymph node metastases. Clin Cancer Res *12*, 2745-2751.

Waki, F., Ando, M., Takashima, A., Yonemori, K., Nokihara, H., Miyake, M., Tateishi, U., Tsuta, K., Shimada, Y., Fujiwara, Y., and Tamura, T. (2009). Prognostic factors and clinical outcomes in patients with leptomeningeal metastasis from solid tumors. J Neurooncol *93*, 205-212.

Wang, W., Epler, J., Salazar, L. G., and Riddell, S. R. (2006). Recognition of breast cancer cells by CD8+ cytotoxic T-cell clones specific for NY-BR-1. Cancer Res *66*, 6826-6833.

Wang, Y., Dong, Q., Miao, Y., Fu, L., Lin, X., and Wang, E. (2013). Clinical significance and biological roles of SPAG9 overexpression in non-small cell lung cancer. Lung Cancer *81*, 266-272.

Wang, Y., Yang, W., Pu, Q., Yang, Y., Ye, S., Ma, Q., Ren, J., Cao, Z., Zhong, G., Zhang, X., *et al.* (2015). The effects and mechanisms of SLC34A2 in tumorigenesis and progression of human non-small cell lung cancer. J Biomed Sci *22*, 52.

Whitaker, H. C., Warren, A. Y., Eeles, R., Kote-Jarai, Z., and Neal, D. E. (2010). The potential value of microseminoprotein-beta as a prostate cancer biomarker and therapeutic target. Prostate *70*, 333-340.

Wilson, C. L., and Matrisian, L. M. (1996). Matrilysin: an epithelial matrix metalloproteinase with potentially novel functions. Int J Biochem Cell Biol *28*, 123-136.

Wilson, E. H., Weninger, W., and Hunter, C. A. (2010). Trafficking of immune cells in the central nervous system. J Clin Invest *120*, 1368-1379.

Yang, Y., and Rosenberg, G. A. (2011). MMP-mediated disruption of claudin-5 in the blood-brain barrier of rat brain after cerebral ischemia. Methods Mol Biol *762*, 333-345.

Yatabe, Y., Dacic, S., Borczuk, A. C., Warth, A., Russell, P. A., Lantuejoul, S., Beasley, M. B., Thunnissen, E., Pelosi, G., Rekhtman, N., *et al.* (2019). Best Practices Recommendations for Diagnostic Immunohistochemistry in Lung Cancer. J Thorac Oncol *14*, 377-407.

Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Trevino, V., Shen, H., Laird, P. W., Levine, D. A., *et al.* (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun *4*, 2612.

Youlden, D. R., Cramb, S. M., and Baade, P. D. (2008). The International Epidemiology of Lung Cancer: geographical distribution and secular trends. J Thorac Oncol *3*, 819-831.

Zafrakas, M., Petschke, B., Donner, A., Fritzsche, F., Kristiansen, G., Knuchel, R., and Dahl, E. (2006). Expression analysis of mammaglobin A (SCGB2A2) and lipophilin B (SCGB1D2) in more than 300 human tumors and matching normal tissues reveals their co-expression in gynecologic malignancies. BMC Cancer *6*, 88.

Zehentner, B. K., and Carter, D. (2004). Mammaglobin: a candidate diagnostic marker for breast cancer. Clin Biochem *37*, 249-257.

Zhang, X. H., Jin, X., Malladi, S., Zou, Y., Wen, Y. H., Brogi, E., Smid, M., Foekens, J. A., and Massague, J. (2013). Selection of bone metastasis seeds by mesenchymal signals in the primary tumor stroma. Cell *154*, 1060-1073.

Zhao, W. H., Xiao, J. Z., Yang, W. Y., Wang, N., Wang, X., Chen, X. P., and Bu, S. (2006). [Relationship between hepatic insulin resistance and the expression of genes involved in hepatic glucose output]. Zhonghua Gan Zang Bing Za Zhi *14*, 45-48.

Zheng, M. M., Li, Y. S., Jiang, B. Y., Tu, H. Y., Tang, W. F., Yang, J. J., Zhang, X. C., Ye, J. Y., Yan, H. H., Su, J., *et al.* (2019). Clinical Utility of Cerebrospinal Fluid Cell-Free DNA as Liquid Biopsy for Leptomeningeal Metastases in ALK-Rearranged NSCLC. J Thorac Oncol *14*, 924-932.

Tables

Table 1. Summary of cell type identity in scRNA-seq results of cancer patient CSF samples.

Patient ID	Diagnostics	Cell selection		Number of sequenced cells	Number of QC filtered cells	T cells	Monocytes	CTCs
		Calcein Blue AM	CD45					
P1-1	LUAD-LM	+	−	168	99	0	0	99
P1-2	LUAD-LM	+	−	360	280	0	0	280
P2	LUAD-LM	+	−	192	130	0	0	130
P4	LUAD-LM	+	−	288	157	0	15	142
P6	LUAD-LM	+	−	480	281	0	0	281
P7	LUAD-LM	+	−	288	205	0	0	205
P3	LUAD-LM	+	N/A	96	50	0	50	0
P8-1	CUP-LM	+	−	480	152	0	0	152
P8-2	CUP-LM	+	−	816	343	28	19	296
Total				**3,168**	**1,697**	**28**	**84**	**1,585**

LUAD: lung adenocarcinoma; CUP: cancer of unknown primary site; LM: leptomeningeal metastases; Calcein Blue AM: labeling dye for live cells selection; CD45: protein tyrosine phosphatase receptor type C, marker for leukocytes; N/A: Not applicable. P1-1 and P1-2 are CSF samples collected from the same patient with a 67-day time interval. Sample P8-2 was collected 158 days after P8-1 (Table S1).

Figures

Figure 1. Isolation of cerebrospinal fluid (CSF) circulating tumor cells (CTCs) and characterization of normal CSF cell composition at single-cell level.

(***A***) CSF-CTC isolation workflow. Patient CSF samples were collected from lung adenocarcinoma (LUAD) leptomeningeal metastases (LM) patients (Table 1). Under CSF cytology (400× magnification; Wright's stain), CTCs have a larger cell diameter (red arrows) than the smaller lymphocytes. Fluorescence-activated cell sorting (FACS) was used to sort individual live cells from LUAD-LM CSF samples, and CD45− cells were subsequently selected for scRNA-seq following the SMART-seq2 protocol. For control purpose, live cells from non-metastatic CSF samples (Normal, N1-N3) were processed using the same pipeline but without cell selection.

(***B***) t-distributed stochastic neighbor embedding (t-SNE) plot of control CSF cells (N1, N2 and N3) and blood cells (Table S1), showing the cell identity and gene expression correlation.

(***C***) High-resolution heat map showing expression of selected marker genes in blood and CSF cells (Leu, Leukocytes; Mon, Monocytes).

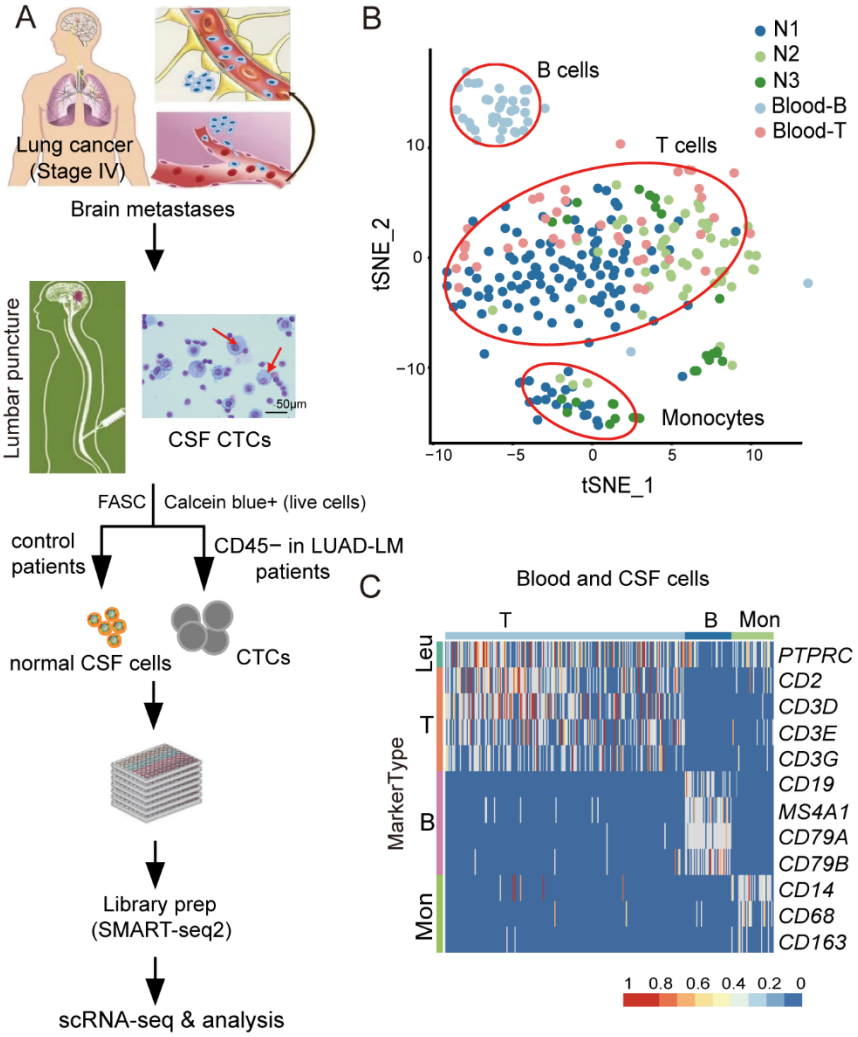


Figure 2. Characterization of LUAD-LM CSF-CTCs using single-cell transcriptome analyses.

(***A***) t-SNE plot of gene expression clustering of three normal CSF samples (N1, N2 and N3) and five LUAD-LM CSF samples (P1, P3, P4, P6 and P7, see Table 1). Clusters are assigned to indicate cell identity and gene expression correlation.

(***B***) Heatmap showing expression levels of selected marker genes in each sample (Leu, Leukocytes; Mon, Monocytes; Epi, Epithelial; Pro, Proliferative).

(***C***) The immune signature of CSF cells quantified by the ImmuneScore computed from the ESTIMATE algorithm, showing significant difference between the leukocyte group (***left***) and the CSF-CTC group (***Right***) ($P < 0.001$, Wilcoxon Rank-Sum test).

(***D***) Volcano plot of significant differential gene expression between the CSF-CTCs (adjusted *P*-value < 0.05) in control and patients. Up-regulated and down-regulated genes are defined using a fold-change cutoff of 2. Gene names are labeled for selected upregulated genes in CSF-CTCs, including important internal reference and marker genes.

(***E***) Significantly enriched KEGG pathway in LUAD CSF-CTCs compared to leukocytes by GSEA (Gene Set Enrichment Analysis) (FDR: 0.05-0.20).

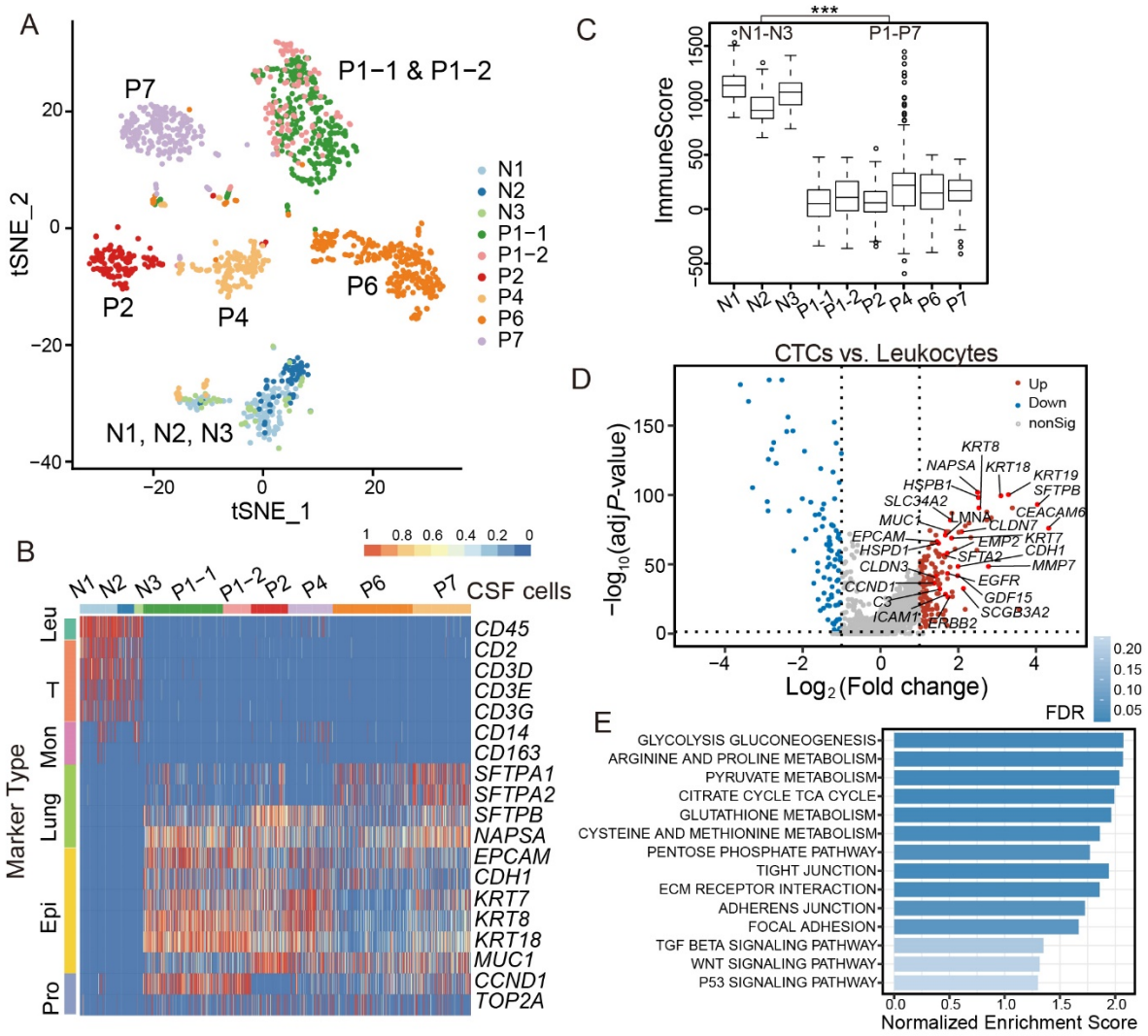


Figure 3. The heterogeneity of LUAD-LM CSF-CTCs among different patients and within individual patient.

(***A***) Locations of leptomeningeal metastases in LUAD-LM patients reveal special heterogeneity. Brain MRI (Magnetic resonance imaging) diagnostic results were shown for patients P1, P2, P4 and P6, demonstrating leptomeningeal enhancement (pointed using red arrows, Table S1).

(***B***) Heatmap showing pairwise correlations between the expression profiles of LUAD-LM patient CSF-CTCs, human non-small cell lung cancer cell line H358, and human lung adenocarcinoma patient-derived xenograft (PDX) cells MBT15 and PT45.

(***C***). ***Top***: degree of heterogeneity among cell measured by mean correlation coefficient within individual samples. ***Bottom***: heterogeneity analysis showing mean correlation coefficients for CTCs within individual CSF samples (intra-patient), among CSF samples (inter-patient), and for cells within individual two PDX samples and H358 cell line (intra-others).

(***D***) Heatmap of enrichment for differentially expressed genes across five LUAD-LM patients. Ordered gene numbers and gene names used in the plot are labeled.

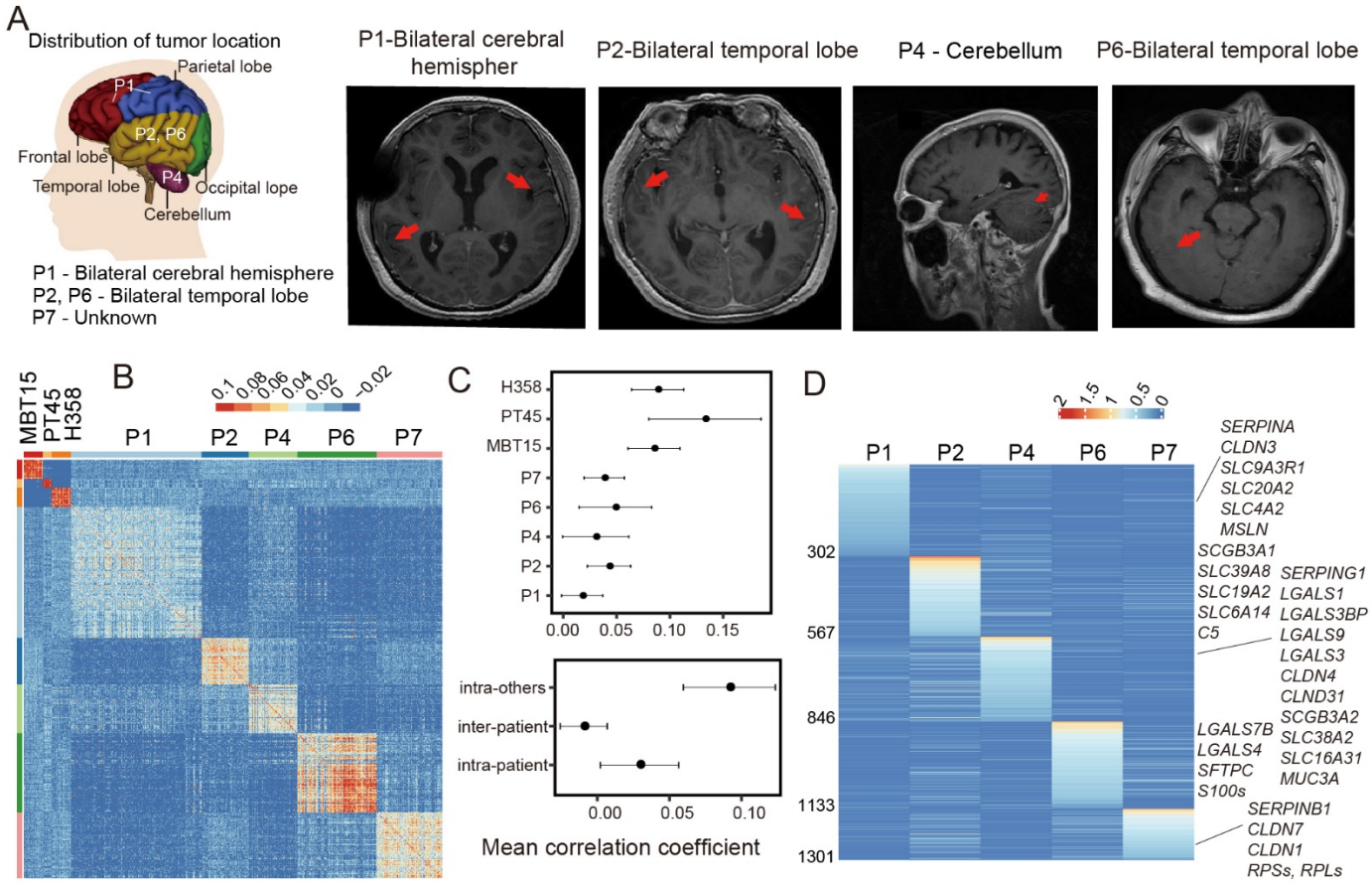


Figure 4. Gene expression profile of cell cycle genes, cancer-testis antigens (CTAs) and genes related to partial EMT in CSF-CTCs.

(***A***) Cell cycle state of individual CSF-CTCs (dots) estimated based on relative expression of G1/S (*x*-axis) and G2/M (*y*-axis) genes in patient P1 and P4. Cells are colored by their inferred cell cycle states (cycling cells: score > 0.2, red; intermediate: 0 < score ≤ 0.2, pink; noncycling cells: score ≤ 0, gray).

(***B***) Unsupervised clustering using GSVA (Gene Set Variation Analysis) score of epithelial genes, mesenchymal/CSC genes and extracellular matrix genes in P1 and P4 CSF-CTCs. Epithelial genes: *CDH1*, *EPCAM*, *KRT18*, *KRT19*, *KRT7*, *KRT8* and *MUC1*; mesenchymal/CSC genes: *FN1*, *VIM* and *CD44*; extracellular matrix genes: *LAMA3*, *LAMA5*, *LAMB2*, *LAMC1*, *ITGA3*, *ITGB4* and *CD47*.

(***C***) Line chart of average normalized expression level (*y*-axis) of epithelial genes (red), mesenchymal/CSC genes (green) and extracellular matrix genes (blue) in P1 and P4. CSF-CTCs were ranked by average normalized expression level of mesenchymal/CSC genes (*x*-axis).

(***D***) Heatmap of single-cell expression profiles of three cancer-testis antigens (*XAGE1B*, *BRDT* and *LY6K*) in CSF-CTCs from five LUAD-LM patients.

(***E***) Boxplot of the number of expressed CTAs (*y*-axis) in CSF-CTCs from five patients (*x*-axis).

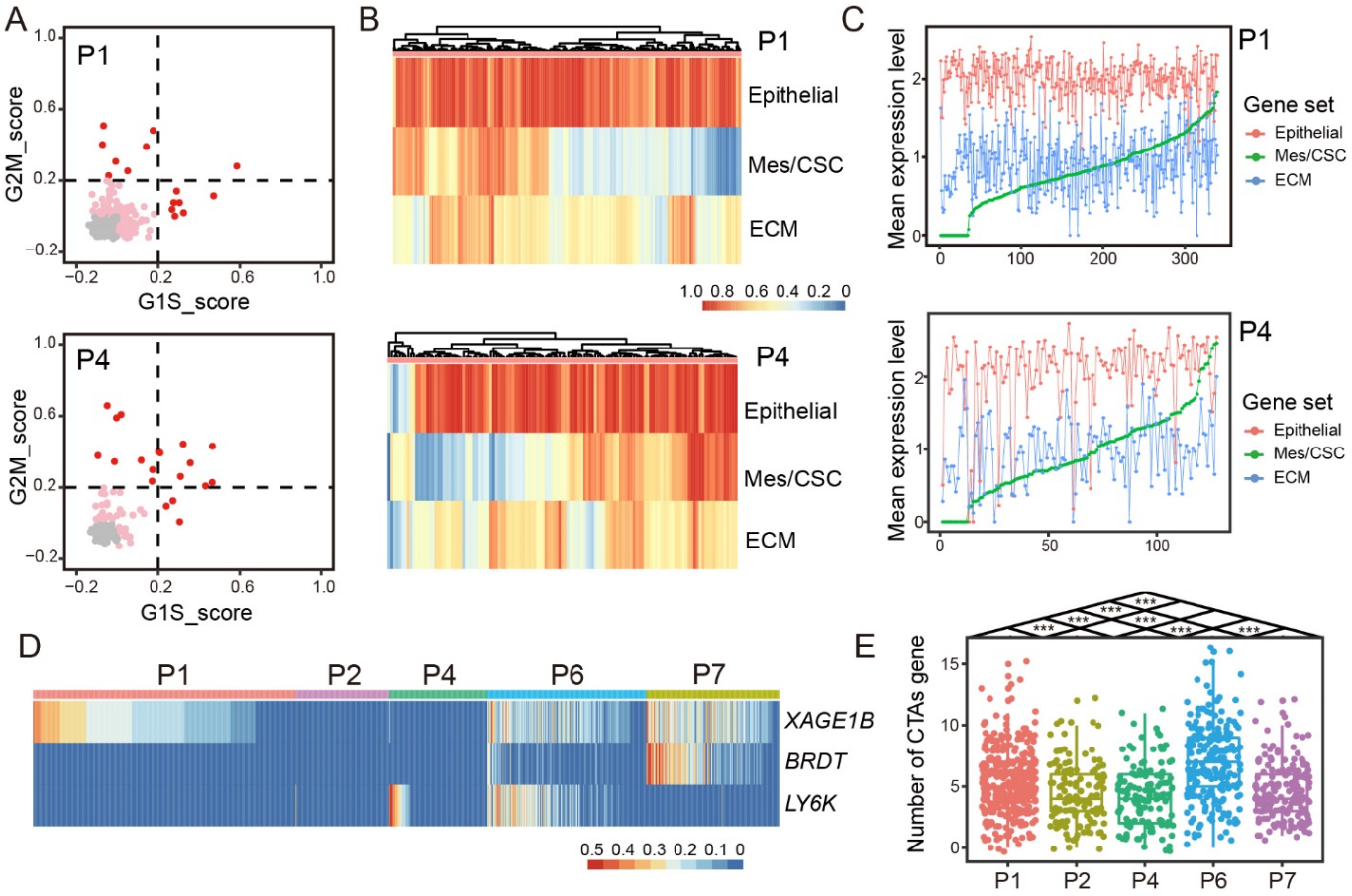
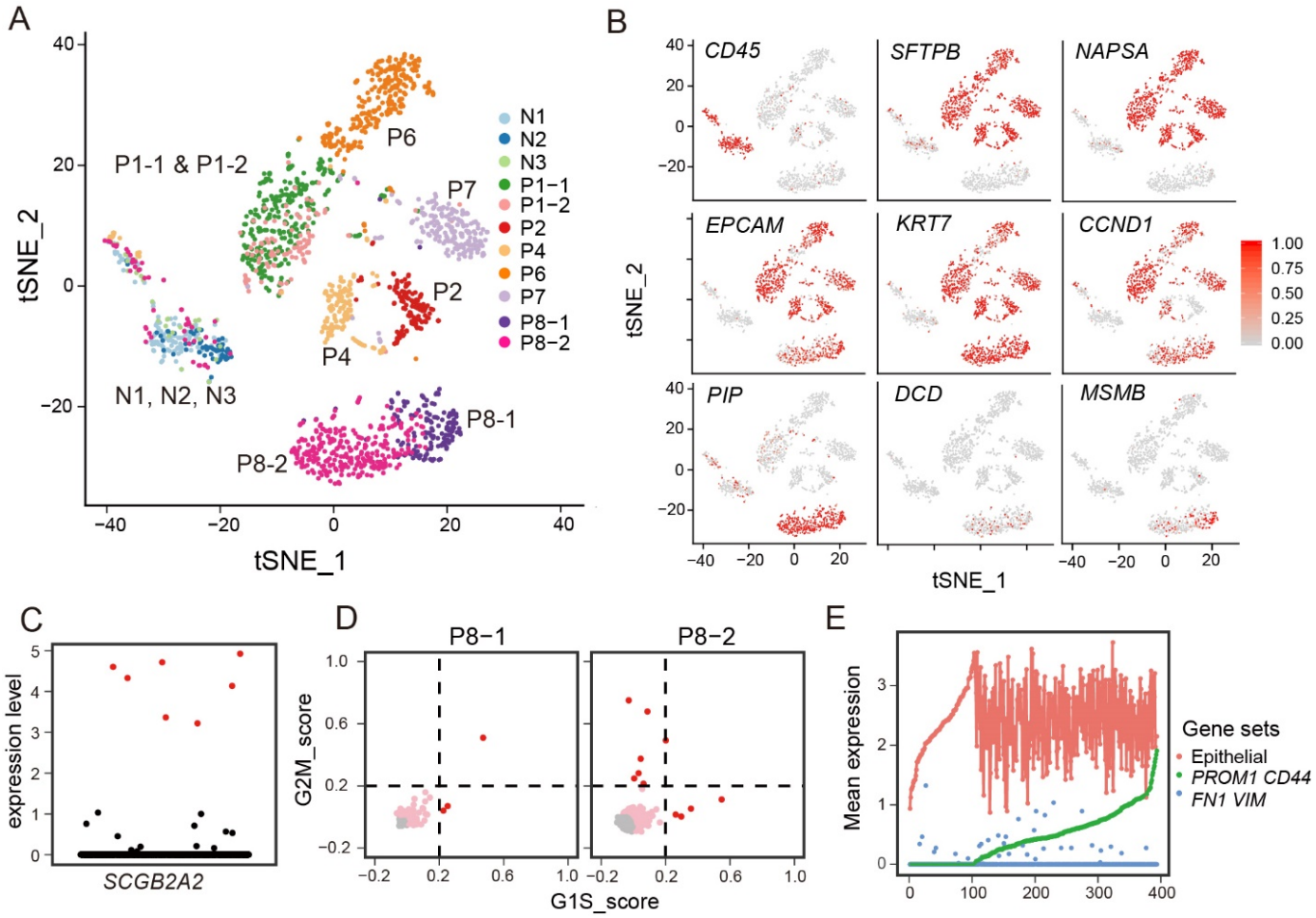


Figure 5. Investigation of a leptomeningeal metastases (LM) case of cancer of unknown primary site (CUP) through scRNA-seq profiling of CSF-CTCs.

(***A***) t-SNE plot of gene expression clustering of patient P8 with normal CSF samples (N1, N2 and N3) and LUAD-LM CSF samples (P1, P2, P4, P6 and P7). P8-1 and P8-2 are two independent CSF samples 6 months apart.

(***B***) Feature plots demonstrating expression of selected genes on the t-SNE plot (Figure 5A). Scaled expression levels are depicted using a red gradient. *CD45*: leukocyte marker. *SFTPB* and *NAPSA*: lung markers. *EPCAM* and *KRT7*: epithelial marker. *CCND1*: proliferating cells marker. *PIP*, *DCD* and *MSMB*: CUP patient P8 specific genes.

(***C***) Normalized expression levels of *SCGB2A2* in all cells in Figure 5A. Seven cells in P8 with high expression of *SCGB2A2* are labeled in red.

(***D***) Cell cycle state of individual CSF-CTCs (dots) inferred from relative expression of G1/S (*x*-axis) and G2/M (*y*-axis) gene sets in P8-1 and P8-2 samples.

(***E***) Average normalized expression level (*y*-axis) of epithelial genes (*EPCAM*, *KRT18*, *KRT19*, *KRT7*, *KRT8*, *MUC1*), mesenchymal genes (*VIM*, *FN1*) and CSC genes (*PROM1*, *CD44*) in P8 cells (*x*-axis).