의학박사 학위논문

Analysis of genetic heterogeneity of tumor-infiltrating immune cells in human cancer by single-cell RNA sequencing

단일세포 전사체 분석을 통한 암미세환경 내 면역세포의 이질성에 관한 연구

2018 년 2 월

서울대학교 대학원

의과학과 의과학전공

엄 혜 현

A thesis of the Doctor's degree

단일세포 전사체 분석을 통한 암미세환경 내 면역세포의 이질성에 관한 연구

Analysis of genetic heterogeneity of tumor-infiltrating immune cells in human cancer by single-cell RNA sequencing

February 2018

The Department of Biomedical Sciences,

Seoul National University

College of Medicine

Hye Hyeon Eum

Analysis of genetic heterogeneity of tumor-infiltrating immune cells in human cancer by single-cell RNA sequencing

by

Hye Hyeon Eum

A thesis submitted to the Department of Biomedical Sciences in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Medical Science at Seoul National University College of Medicine

December 2017

Approved by Thesis Committee:

Professor _____ **Chairman**

Professor _____ **Vice chairman**

Professor _____

Professor _____

Professor _____

ABSTRACT

Introduction: Tumor environment is established by various components including the malignant, stromal and immune cells. Single-cell transcriptome profiling of tumor samples allows the dissection of heterogeneous tumor cells and the neighboring stromal or immune cells. Precise characterization of the infiltrating immune cells may provide clues for novel immunotherapy strategies.

Methods: A total of 515 individual cells from 11 breast cancer patients and 162 cells from four advanced gastric cancer (AGC) patients were analyzed by single-cell RNA sequencing (RNA-seq). Reference single-cell transcriptomes for M1-type or M2-type macrophages were generated from normal blood-derived monocytes after *in vitro* differentiation.

Results: Copy number alteration patterns inferred from the single-cell RNA-seq data separated tumor cells and non-tumor cells. Most of the non-tumor cells were immune cells. In breast cancer*, three distinct immune cell clusters of T lymphocytes, B lymphocytes, and macrophages were identified. T lymphocytes displayed immunosuppressive characteristics with a regulatory or exhausted phenotype. B lymphocytes were divided into two subgroups, the anti-apoptotic naïve/memory cell group and the highly proliferative B cell group.

In AGC, all of the detected immune cells were tumor-associated macrophages (TAMs). When compared to the reference transcriptomes of the M1 or M2 macrophages, an M2-biased tendency was observed in the AGC TAMs with a heterogeneous level of polarization. In comparison, TAMs originating from breast cancer or colorectal cancer showed both M1-biased and M2-biased cells.

Conclusions: This study demonstrates the power of single-cell RNA sequencing for the characterization of tumor-infiltrating immune cells to develop immunotherapeutic strategies. Single-cell transcriptome analysis subdivides the microenvironmental immune cells by cell types and their properties with heterogeneous levels of pathway activation. High resolution profiles can provide a new view on targeting the exhausted tumor-infiltrating lymphocytes or transforming the anti-inflammatory M2-type TAM to the pro-inflammatory M1-type.

* This work is based on published article in *Nature Communications* (1).

Keywords: Single-cell, Breast cancer, Advanced gastric cancer, Heterogeneity, Tumor microenvironment, Tumor-associated macrophages

Student number: 2012-21802

CONTENTS

LIST OF FIGURES

Part I

Part II

LIST OF TABLES

Part I

Part II

LIST OF ABBREVIATIONS

RNA-seq : RNA sequencing

TIL : tumor-infiltrating lymphocytes

TAM : tumor-associated macrophage

BC : breast cancer

ER : estrogen receptor

PR : progesterone receptor

HER2 : human epidermal growth factor receptor 2

TNBC : triple-negative breast cancer

AGC : advanced gastric cancer

CRC : colorectal cancer

TPM : transcript per million

CNA : copy number alteration

GSVA : gene set variation analysis

NMF : non-negative factorization

ROC : receiver operating characteristic

LRT : likelihood-ratio test

PCA : principle component analysis

SNN : shared nearest neighbor

tSNE : t-distributed stochastic neighbor embedding

DEG : differentially expressed gene

INTRODUCTION

Gene expression profiling of bulk tumors reflects the features of non-tumor compartments, which are characterized by an admixture of stromal, immune, and endothelial cell infiltration, not only tumor cells. These admixtures form the tumor microenvironment and play a critical role in tumor initiation, progression, and treatment resistance. Microenvironmental gene expression signatures may present prognostic values independent of the intrinsic tumor subtype (2-4). Immune cell infiltrates are composed of cells from multiple lineages that may play pleiotropic roles in tumor immunity. Tumor-associated macrophages (TAMs) often promote tumor progression and metastasis, whereas CD8+ cytotoxic T cells and CD4+ Th1 cells elicit antitumor immunity and suppress tumor growth (5). Furthermore, T cells with regulatory or exhausted phenotype are associated with failure in antitumor immunity. A subset of B cells was proposed to promote tumor progression by affecting diverse cell types including T cells and TAMs (6). However, the presence of a large number of B cells in the tumor region is associated with a good prognosis (7). Altogether, the tumor microenvironment is formed through interactions between these variable cellular components and through communication with tumor cells. Whole genomic profile of the tumor microenvironment using bulk sample has limitations in finding cell type specific characters due to mixed cell composition and their complex interaction.

Immunotherapy has become a promising strategy for cancer treatment since knowledge of the microenvironment has become more established. Targeting suppressor T cells is one certified method of immunotherapy that may revitalize a suppressed immune system. Typical molecules that could change cytotoxic CD8+ T cells to the exhausted phenotype are PD-1 (PDCD1), which is found in T cells, and its ligand PD-L1 (CD274), which is expressed in antigen-presenting cells (8). In particular, blocking antibodies against PD-1 or PD-L1 were shown to have clinical efficacy in the treatment of melanoma and lung cancer (9) and has been applied to cancer types with the approval of the FDA. Many studies on the prognostic and therapeutic application of TAM are ongoing (10-15). Based on the anti-inflammatory function of M2-type macrophages, common methods used to investigate the impact of TAM invasion are the comparison of the M1/M2 ratio and the percentage of M2-type macrophages by detecting standard markers such as HLA-DR or iNOS for M1-type and CD163 for M2-type macrophages (14, 15). However, marker-based approaches cannot reflect the polarization level of the macrophages or distinguish macrophages from other cells which also express same markers.

Since Navin and his colleagues have proposed single-cell genome sequencing as a new approach for revealing the heterogeneity of tumors (16), many researchers have applied single-cell technology to cancer research (17, 18). Among the genomic, transcriptomic, proteomic, and epigenomic platforms, the most common method currently is the single-cell RNA sequencing. Knowing the level of transcriptomic heterogeneity in the tumor

and the precise characterization of tumor and microenvironmental gene expression may help identify better molecular targets for prognosis and treatment (19). The characterization of tumor infiltrating immune cells with a single-cell RNA-seq may also provide a better strategy for overcoming immunosuppression and activating spontaneous immune surveillance (20).

The general treatment procedure for breast cancer depends on the tumor subtype and the subtype classification criteria, which are well established. The traditional method for breast cancer subtyping is a histological examination of estrogen receptors (ER), progesterone receptors (PR), and human epidermal growth factor receptor 2 (HER2). Gene expression-based molecular subtyping (21, 22) and the application of molecular subtypes to RNA-seq data (23) support accurate treatment decisions. Since ER and HER2 are the clear markers and also the incontrovertible cause of cancer development, a treatment method that targets ER or HER2 pathways is a standard regimen for the ER-positive or HER2-positive subtypes, and is usually successful. In the case of triple-negative breast cancer (TNBC) type, however, there are no specific marker molecules for a targeted therapy even in expression profiles. Moreover, approximately 20% of ER-positive tumors are not treated with ER pathway-targeting endocrine therapy. Novel strategies for these cases may be developed through the investigation of a heterogeneous tumor's environment.

There are many environmental and genetic factors that influence the development of gastric cancer, including Helicobacter pylori (H. pylori) infection, smoking, diet (particularly one that is salt-rich), Epstein-Barr virus (EBV) infection, and various genetic and epigenetic changes that are

associated with the incidence of tumors (24-28). At an early stage, there are no certain symptoms except stomach discomfort, but as the cancer progresses, some evident symptoms that may appear include weight loss for an unknown reason, stomach pain, jaundice, and ascites. The considerable influence of environmental factors and the difficulty of recognizing symptoms often lead to a late diagnosis. For these reasons, gastric cancer is often diagnosed at an advanced stage, and is therefore very difficult to cure. Peritoneal ascites is a typical symptom of stage IV advanced gastric cancer (AGC). Malignant ascites contains a large amount of macrophages and peritoneum-derived mesothelial cells, as well as malignant and tumor-associated immune cells that originated from the primary tumor or peritoneal metastasis (29-31).

To investigate heterogeneous characteristics of immune cells in the breast cancer microenvironment, 515 single cells from 11 patients with different breast cancer subtypes were analyzed in Part I. Single cell isolates from individual tumor tissues contained tumor and non-tumor microenvironmental cells. The immune populations were clearly disclosed; however, a detailed characterization of TAM was infeasible due to its low occurrence. In Part II, 72 macrophage single-cell profiles were obtained from among 162 cells taken from four advanced gastric cancer patients. In addition, a total of 142 transcriptomes for M1 or M2 macrophages were constructed from two healthy donors. Reference M1 and M2 cells were used to reveal the polarization level of the macrophage cells objectively and detect M1-polarized or M2-polarized cells within TAMs.

This study demonstrates the need for single-cell profiling to discover tumor-infiltrating lymphocytes (TILs) or TAMs with tumor-promoting features among heterogeneous cells. This approach could provide a clue to more accurate and efficient therapeutic methods.

MATERIALS AND METHODS

1. Patients and tumor-derived specimens

All patients and healthy donors in this study agreed to provide biospecimens through a consent form approved by the Institutional Review Board of Seoul National University Hospital and Samsung Medical Center (Institutional Review Board no. 1207-119-420, 2015-12-094-003 and 2016-04-107-003). Primary breast cancer tissues were obtained from 11 patients diagnosed with invasive ductal carcinoma (IDC) and two metastatic lymph nodes were also collected from patient BC03 and BC07 (BC03LN, BC07LN). Ten of 11 patients had no any treatment prior to surgery but one patient (BC05) received neoadjuvant chemotherapy and Herceptin treatment. Regarding transcriptomic changes resulting from the drug response, the BC05 sample was considered as an outlier and excluded after cell type definition. Molecular subtypes of breast tumors were predicted by transcriptomic profiles using the R package genefu (33). Peritoneal ascites from four AGC patients and one cerebrospinal fluid of patient AGC04 (AGC04CSF) were separately collected for the treatment purpose. These five specimens derived from AGC patients were utilized for single-cell RNA sequencing.

2. Single cell isolation and cDNA amplification

To prepare single-cell suspensions, the specimens of breast cancer tissues or lymph node metastases were dissociated for 2 hours immediately after surgery using both of enzymatic and mechanical methods. Suspending cells in

ascites or cerebrospinal fluid from advanced gastric cancer patients were collected by centrifuge. Dead cells in the single-cell isolates were removed by Ficoll gradient separation using Ficoll-Paque^{TM} PLUS (17-1440-02, GE Healthcare, Uppsala, Sweden), and finally 50,000 live cells were loaded onto an individual integrated fluidic circuit mRNA sequencing chip in the C1^{TM} Single-Cell Auto Prep System (100-5760, Fluidigm, San Francisco, CA, USA). Applied chip type was determined by size distribution and average cell size of each sample. The 17–25-μm targeting chip was used only for AGC03 and AGC04CSF samples, and rest of samples were loaded onto 10–17-μm targeting chip.

After loading cells, single-cell capture was microscopically examined before cell lysis step. The single-cell amplification step including cell lysis, cDNA synthesis, and amplification was performed using the SMARTer Ultra Low RNA Kit (634936, Clontech, Mountain View, CA, USA) as the manufacturer's recommendation. Following the manufacturer's instructions, I added RNA spike-ins 1, 4, and 7 from ArrayControl^{TM} RNA Spikes (AM1780, ThermoFisher, Waltham, MA, USA) to the lysis mix as experimental controls and I utilized these spike-ins to examine constant amplification at a quality check analysis (Figure 1-2A). Amplified cDNAs from single-cells were quantified by a Qubit® 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and qualified using a 2100 Bioanalyzer system (Agilent Technologies, Santa Clara, CA, USA) with high-sensitivity reagents. To run matched bulk tumor samples, total RNAs were extracted from ~1 × 10^5 cells or tumor tissues using the RNeasy Plus Micro kit (74034, Qiagen, Hilden, Germany)

and 10 ng of total RNA was used. Matched bulk profiles were generated under the same conditions as used for single cells from the cell lysis step.

3. RNA-seq and data processing

Successfully amplified 901 single-cell cDNAs (579 of breast cancer cDNAs, 180 of AGC cDNAs, and 142 of macrophage cDNAs) were respectively subjected to RNA sequencing in Part I and Part II. The sequencing libraries were constructed using the Nextera XT DNA Sample Prep Kit (FC-131-1024, Illumina) according to the manufacturer's instructions, and then the constructed libraries were sequenced using the HiSeq2500 system in 100-bp paired-end mode. To assess the expression values of array control RNA spike-ins, reference sequences and the corresponding annotations were generated by merging three control RNA spike-ins (ThermoFisher) with the human genome reference sequences (hg19) and the GENCODE 19 annotations. The RNA reads were then aligned to the reference sequences using the 2-pass mode of STAR_2.4.0b (default parameters) (34), and relative gene expression was quantified as transcript per million (TPM) using RSEM v1.2.17 (default parameters) (35). Isoform expression levels for each gene were summed to derive the TPM values. Quality control assessment of aligned single-cell RNA-seq reads was performed using RNA-SeQC (36). To remove cells with low-quality sequencing values, four filtering criteria were applied; 1) number of total reads, 2) mapping rate, 3) number of detected genes, and 4) portion of intergenic region.

4. Estimation of gene expression level

In breast cancer study, genes with low expression values were trimmed out by the following steps. First, TPM values less than one were considered unreliable and substituted with zero. Second, TPM values were log2-transformed after adding a value of one. Third, genes expressed in less than 10% of all tumor groups were removed. In total, 515 single cells and 17,779 genes passed the QC criteria. The filtered 17,779 genes were also used for bulk tumor analyses.

For transcriptome analysis, expression data were mean centered by subtracting the average log2(TPM+1) value for each gene with the following exceptions: for RNA spike-in analysis, comparisons between RNA-seq and qPCR results, measurement of intra-tumoral correlations, detection of chromosomal expression patterns, comparisons with immuno-fluorescence staining results, comparisons of immune marker expression, and application of self-normalizing tools such as geneset variation analysis (GSVA). As an estimate of the sensitivity and reproducibility of single-cell RNA sequencing, I obtained consistent log2(TPM+1) ratios for RNA spike-in 1 (12,200 transcripts) and RNA spike-in 4 (912 transcripts) (Figure 1-2A). RNA spike-in 7 with an estimated 62 transcript input was not detected.

To certainly reduce effects of outlier and focus on remarkable features, TPM values were divided by 10 before log2-transformation in Part II study as Tirosh did (18). After copy number estimation from RNA-seq data for identifying tumor and non-tumor cells, eighteen cells expressed less than

1,000 genes were eliminated in further analysis. For transcriptome analysis, only the genes of log2(TPM/10+1) > 1 expressed in more than 10 cells each time was used. Except the cases of running self-normalizing GSVA or being expressed has its own meaning, expression values were rescaled by mean-centering by each gene. I decided that being expressed has a meaning itself when inferring copy number alterations (CNAs) or comparing marker gene expression.

5. Quantitative PCR for validation of RNA-seq

A small amount quantitative PCR (qPCR) was performed with the 192.24 Dynamic Array^{TM} integrated fluidic circuit using DELTA gene assay (PN100-3035, Fluidigm) to validate the expression level of single-cell RNA-seq data. Six breast cancer samples with diverse subtypes were selected, then six bulk cDNAs and corresponding 185 single-cell cDNAs were used in qPCR. Twenty-two representative genes including breast cancer subtype-specific markers (ESR1, PGR, and ERBB2) and cell-type specific markers (EPCAM, PTPRC, CDH1) were selected and two housekeeping genes (GAPDH and ACTB) were added. All primers were designed using the manufacture's recommended software D3 assay design (PN 100-6812, Fluidigm) and are listed in Table 1-2.

Before comparing qPCR and RNA-seq data, the Ct values of qPCR were converted to scales comparable to log2(TPM+1) of RNA-seq. Ct values were converted to negative values and genes with too low expression (-Ct < -20) were adjusted to -20. The inter-relations of two platforms were evaluated by

Pearson's correlation, Spearman's rank order correlation, and linear regression analysis.

6. Inference of chromosomal expression patterns

Chromosomal expression pattern gives highly comparable information to CNA (17). Cancer cells tend to exhibit distinct chromosomal expression pattern when compared to non-carcinoma cells, thus I used this approach to distinguish cancer cells from non-tumor cells. To prepare normal (baseline) expression level of breast or stomach tissues, firstly normal tissue expression profiles for breast or stomach were downloaded from GTEx v6 datasets (http://www.gtexportal.org/). Second, downloaded normal expression data was transformed to the corresponding scale as our data. Third, the average gene expressions of each tissue types were calculated for the overlapped genes with our filtered single cell data.

To estimate gene expression patterns in the genomic position, the genes were sorted by their transcriptional start position. For the breast cancer analysis, the Z-scores of our breast cancer single-cell data for each gene were calculated by normalizing with the averaged expression profile of normal breast tissues. Then the inferred CNA patterns were estimated by the moving averages of 150 genes and centered values across genes per chromosome were used. For the advanced gastric cancer analysis, I applied the advanced process to remove an outlier effect (18). First, the expression data of our single-cell and GTEx normal tissue transcriptomes were separately normalized by Z-scoring for each gene. Relative expression values were restricted to a range of

-3 to 3, then the moving averages of 150 genes were calculated and centered across genes. Lastly, the average of the estimated CNA values for the normal expression profiles were subtracted from the estimated CNA values of single-cells.

7. Immune cell type-specific gene expression profiling

Three immune cell subgroups were identified in breast tumor micro-environment by non-negative factorization (NMF) clustering (37) from 175 non-tumor cells using 412 genes annotated in 11 non-overlapping immune cell types (38) (Table 1-3). To characterize the three immune cell clusters, the Receiver operating characteristic (ROC) test and likelihood-ratio test (LRT) based on zero-inflated data were performed using R package Seurat (39). Then, genes with a fold-change > 2 and an AUC > 0.7 were obtained as cell type-specific genes (LRT $P<0.05$) (Table 1-4). Gene ontology terms significantly enriched in cell type-specific genes were annotated by DAVID 6.7 (https://david.ncifcrf.gov/) with a default option. To further characterize T or B cells by functional status, geneset variation analysis (GSVA) analysis was performed with selected gene sets from the literature (Table 1-5).

8. Immunofluorescence staining

Immunofluorescence staining was carried out to assess the presence of tumor-infiltrating T or B cells in breast tumor tissues. T lymphocytes were double-stained with anti-CD3 (1:200; MA5-12577, Thermo Fisher, Waltham,

MA, USA) and anti-MARK3 (1:100; PA5-29328, Thermo Fisher) antibodies in the FFPE slides. B lymphocytes were double-stained with anti-CD20 (1:200; MA5-13141, Thermo Fisher) and anti-PRPSAP2 (1:50; PA5-31237, Thermo Fisher) antibodies. Alexa488 labeled-anti-mouse and Alexa568 labeled-anti-rabbit antibodies (1:50; Invitrogen) were used for double immunofluorescence with DAPI counterstaining. The numbers of CD3+ or CD20+ cells were assessed as average counts in three 0.125mm^2 areas with maximal positive staining.

9. Gene set variation analysis

To estimate pathway activation levels or evaluate gene expression signatures with a gene set level, a gene set variation analysis (GSVA) was performed using the R package gsva (40). Prior to GSVA analysis, over-representation analysis was performed for all gene sets by the hypergeometric test and reliable gene sets with p-value < 0.05 were used.

10. Construction of M1/M2 reference single-cell profiles

Fresh PBMCs of two healthy donors were separated from whole blood samples by Ficoll-Paque^{TM}. Then only CD14+ monocytes were selected out using MACS human CD14 MicroBeads (130-050-201, Miltenyi Biotec GmbH, Bergisch Gladbach, Germany), Pre-Separation Filters (130-041-407, Miltenyi), and LS columns (130-041-401, Miltenyi) as following the manufacture's recommend. To induce M0 macrophages, firstly isolated

CD14+ monocytes were seeded into FBS coated 24-well plate with a density of 1×10^5 cells/cm^2. Seeded cells were cultured 7 days in RPMI1640 media supplied with 20% of FBS and 100ng/ml of M-CSF (574802, Biolegend, San Diego, CA, USA). At day 7, M-CSF containing media were removed and appropriate stimulating media with 100ng/ml LPS (L4524, Sigma-aldrich, St. Louis, MO, USA) and 20ng/ml IFN-γ (570202, Biolegend) for M1 or 20ng/ml IL-4 (574002, Biolegend) and 20ng/ml IL-10 (571002, Biolegend) for M2 were supplied. After additional 48hr of culture, cells were collected by smooth scrapping. A part of cells were triple-stained with PE/Cy7 anti-human CD14 antibody (325618, Biolegend), FITC anti-human CD80 antibody(305206, Biolegend), and PE anti-human CD163 antibody (333606, Biolegend) for fluorescence-activated cell sorting (FACS) analysis. The remaining cells were separately loaded onto the 17–25-μm chips of the C1 system then performed single-cell RNA-sequencing.

11. Scoring of M2-like polarization

To measure the similarity with M1-type or M2-type macrophages in each cell, I extracted M1-specific or M2-specific gene lists by LRT for zero-inflated data using the constructed M1/M2 single-cell data. Genes with fold-change > 2 and p-value of LRT < 0.05 were obtained from each donor and then overlapped genes were considered as core differentially expressed genes (DEGs). The score of each gene set was calculated by GSVA analysis and the M2-like polarization level (M2-like score) was expressed by subtracting the M1 score from the M2 score. Gene sets of M1, M2, or total macrophages

from bulk tissue data were collected from the published data (Table 1-3 and Table 1-5A) (38,41). Cell ordering was performed to visualize the biased polarization of macrophages in various tumor types using the M1/M2 cell specific genes by DDRTree algorithm (42) in R package Monocle 2.0 (43).

RESUTLS (PART-I)

Separation and identification of immune cells from tumors by single-cell RNA sequencing in breast cancer

1-1. Pathologic profiles of patients for single-cell analysis

Selected 11 patients represent the four subtypes of breast cancer: luminal A, luminal B, HER2, and triple negative. All but one of the surgical samples were obtained from chemotherapy-naïve patients, and the markers for subtyping were validated by pathological examination as ER-positive (BC01 and BC02; luminal A), ER/HER2-positive (BC03; luminal B), HER2-positive (BC04, BC05, and BC06; HER2), and triple-negative (BC07-BC11; TNBC) invasive ductal carcinoma (Table 1-1). Regional metastatic lymph nodes were collected from the luminal B (BC03LN) sample and a triple negative breast cancer (BC07LN) sample.

Each tumor tissue manifested a diversely mixed cell composition with a differential level of immune cell infiltration (Figure 1-1). Luminal A type (BC01 and BC02) tumors were primarily enriched with tumor cells, whereas most TNBC type (BC07-11) tumors showed extensive immune cell infiltration (3) (Figure 1-1A). The high proportion of non-tumor components, including large numbers of tumor-infiltrating lymphocytes (TILs), demonstrates the need for single cell analysis to generate an accurate understanding of tumors and their microenvironments.

1-2. Reliability of single-cell RNA sequencing

To reproduce the tumor environment using single-cell sequencing, single cells were isolated using microfluidic chips (44) without prior cell-type selection. After filtering out unreliable, low-quality cells, RNA-seq data

generated from 515 single cells (containing 5.8 ± 1.3 million reads) was analyzed. Detection of constant ratios of two spiked-in RNAs assured the quality and consistency of all single-cell RNA-seq experiments (Figure 1-2A). Quantitative PCR analysis of the expression of 24 selected genes supported the data from single-cell RNA-seq (Figure 1-2B and Table 1-2).

Tissue isolates was highly reflective of tumor tissue (Figure 1-2C). Comparisons between the averages of single cells and corresponding pooled samples (Figure 1-2D) demonstrated partial but significant correlations (Pearson's r 0.16-0.63 with average 0.47, $p < 0.001$). Multiple regression analyses of the transcriptomes of different sized pools of single cells to those of bulk tumors (Figure 1-2E) provided a better representation of the tumor population with an increasing number of single cells. Altogether, single-cell RNA-seq could illustrate a significant portion of the tumor entity, yet tumor components were lost during the single cell isolation or sequencing processes.

1-3. Separation of tumor and tumor-associated normal cells

Because our RNA-seq data was produced without any prior selection of tumor cells, I speculated that non-tumor cells might be captured together with tumor cells. Bulk profiling using tumor tissue reflects the characteristics of all components of the tumor microenvironment, including tumor cells and surrounding microenvironmental cells. When profiling in single-cell analysis, this signature mixing problem is solved and the signatures of specific cell type

could be separated. In previous report, Patel and his colleagues utilized single-cell RNA-seq to distinguish malignant and non-malignant cell in glioblastoma (17). They aligned the gene expression profile along to the genomic position and estimate copy number alterations (CNAs) using moving averages of the gene expression levels. I refer this inferred CNA pattern to chromosomal expression pattern to explain intuitively in this thesis, and used to separate the immune cells from the breast tumor cells and other mixed population.

Unsupervised hierarchical clustering was performed using chromosomal expression patterns of 515 single cells and 11 matched bulk tumor tissues (Figure 1-3A). Total ten tumor-specific clusters from each patient and one large non-tumor cluster were identified. The patient-specific tumor cell clusters showed distinct chromosomal gene expression patterns, whereas non-tumor cell cluster consisted of multiple patient cells and showed no clear patterns. The separation of the two large groups (tumor group vs. non-tumor group) was reconfirmed by the unsupervised principal component analysis (PCA) using the whole transcriptomic profiles (Figure 1-3B). The most principal component (PC1) was greatly affected by the number of detected genes, and the next major component (PC2) may be influenced by genomic instability.

The detected non-tumor population potentially represents stromal cells — such as fibroblasts, adipocytes, or endothelial cells — or diverse immune cells, according to the histopathological examinations of the tumor tissues (Figure 1-1). The expression level of representative marker genes (Figure 1-3C) was compared, and as expected, tumor cells classified using inferred CNAs

expressed numerous epithelial cell differentiation markers (KRT9, CDH1, and EPCAM), which non-tumor cells did not do. Instead, the majority of non-tumor cells may have been immune cells since they expressed immune markers such as PTPRC and IL2RG. Only a few cells showed a fibroblast phenotype through the expression of fibroblast markers (HTRA1, FBN1, and FAP). After the removal of latent fibroblasts and unclassified cells, it is estimated that 175 tumor-associated immune cells were collected out of the 515 single cells.

1-4. Immune cell populations identified in tumor microenvironment

Next, these putative 175 immune cells were further subdivided into three groups using non-negative matrix factorization (NMF) clustering with immune cell type-specific gene sets (38) (Figure 1-4 and Table 1-3). Of the results from k = 2 to k = 9, clustering into three clusters showed the clearest segmentation and the highest cophenetic coefficient value and the average silhouette width (Figure 1-4B and C). These results suggesting an optimal cluster of three.

To investigate the identity of each cluster, cluster-specific genes were extracted by receiver operating characteristics (ROC) test and likelihood-ratio test (LRT) based on zero-inflated data (Figure 1-5A and Table 1-4). Gene ontology terms from the extracted gene sets characterized the identity of each cluster (Figure 1-5B). The largest group expressed immunoglobulins and B

cell-specific transcriptional factors, and many came from the tumor-infiltrated lymph nodes (Cluster 1/B cells; Figure 1-5B and Table 1-4A). The second group expressed T-cell receptors and T cell-specific markers, most of which were captured from primary tumor tissues (Cluster 2/T cells; Figure 1-5B and Table 1-4B). The third group also came from the primary tumor tissues and expressed markers for tissue macrophages (Cluster 3/Macrophages; Figure 1-5B and Table 1-4C).

The presence of tumor-infiltrating T and B lymphocytes was also assessed by immunostaining in tumor tissues with anti-CD3 (d+g+e) or anti-CD20 antibodies (Figure 1-6A), which showed significant correlations with gene expression in the bulk tumor samples ($p < 0.05$) (Figure 1-6B). Among 10 tumor tissues with T lymphocyte-specific gene expression and immunostaining, individual T lymphocyte capture was successful in four tissue isolates (Figure 1-6B). Individual B lymphocyte capture was successful in four out of seven tissues (Figure 1-6B). The results of additional marker staining for T and B lymphocytes were consistent with the single-cell RNA-seq results (Figure 1-6C and D). These data support the validity of gene expression profiling for cell type specification, but also implicate limitations in single cell isolation from breast cancer tissues.

1-5. Heterogeneity in tumor-infiltrating immune cells

Further investigation was carried out using geneset variation analysis (GSVA) to clarify the phenotype of each cell (Figure 1-7). The expression level of immune signatures and marker genes reconfirmed the identity of each

cluster (Figure 1-7A). Comparing the activation levels of pathways associated with B cell functions, two subclasses of B lymphocytes were identified (Figure 1-7B and Table 1-5B). One group, mostly from TNBC tumors (BC07, BC07LN, and BC09), showed naïve/memory B cell gene expression signatures. Another group, mostly derived from the BC03 (luminal B tumor) lymph node, showed proliferation signatures associated with germinal center (GC) centroblasts/centrocytes (45). The formation of GC is an important feature of the antigen-activated immune response.

In the tumor-infiltrating T cells, I analyzed gene expression signatures for T cell activation and functional status (Figure 1-7C and Table 1-5C). T lymphocytes in four patient tumors manifested distinct patterns for naïve, costimulatory, regulatory, exhaustion, and cytotoxicity expression signatures (8,46,47). The luminal B type tumor (BC03) had T lymphocytes with naïve/early costimulatory signatures in the primary tumor sites and T lymphocytes with more costimulatory molecule expression in the lymph nodes. One HER2+ tumor and two TNBC tumors (BC04 and BC07) were populated by T lymphocytes with the expression of regulatory T cell markers including IL2RA (also known as CD25) (47). The third TNBC tumor (BC09) had two types of T lymphocytes, one with a predominant exhaustion signature and another exhibiting both exhaustion and cytotoxicity signatures. T cells with a high exhaustion signature are targets of immune checkpoint blockade in clinical oncology. However, PD-1 expression was modest in our dataset, but expression of alternative inhibitory receptors such as TIGIT and LAG3(48) was frequently detected.

The third immune cell group expressed transcripts for the monocyte/macrophage markers CD14 and CD68 (38) as well as phagocytic enzymes associated with macrophage function, suggestive of TAMs (Figure 1-7A and Table 1-4C) (10). The cluster-specific genes were significantly enriched for genes involved in inflammation and defense mechanisms (Figure 1-5B). Previous reports suggest that TAMs may show an immunosuppressive M2 signature, which promotes tumorigenesis by suppressing immune surveillance and inducing angiogenesis, rather than the activating M1 type signature (41). Indeed, the putative TAM populations expressed many M2-type genes (Figure 1-7A and Table 1-5A) such as CD163, MS4A6A, and TGFBI (41), in addition to genes known to promote tumor progression and angiogenesis such as PLAUR (5) and IL-8 (49). Collectively, both innate and adaptive immune cell populations in the breast cancer samples displayed immune-suppressive gene expression characteristics.

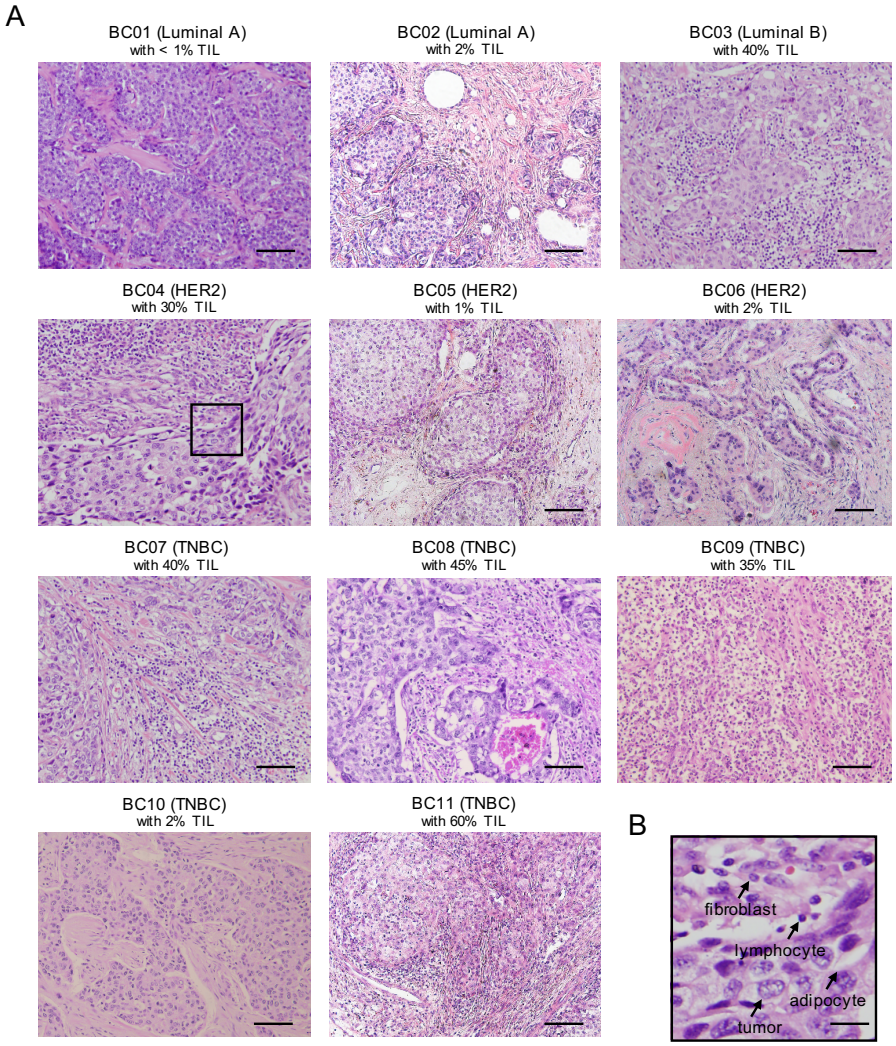

Figure 1-1. Various immune infiltrating levels in breast tumor tissues. (A) Hematoxylin and eosin staining on FFPE slides. Microscopic findings indicated tumor and non-tumor cells, including tumor-infiltrating lymphocytes (TIL, 1-60%). Most of the TNBC tumors except BC10 were heavily infiltrated with lymphocytes whereas luminal A tumors showed enrichment with tumor cells. Scale bar, 100μm. **(B)** A part of the tumor tissue in (A) is magnified to show non-neoplastic cellular components as a representative. Scale bar, 25μm.

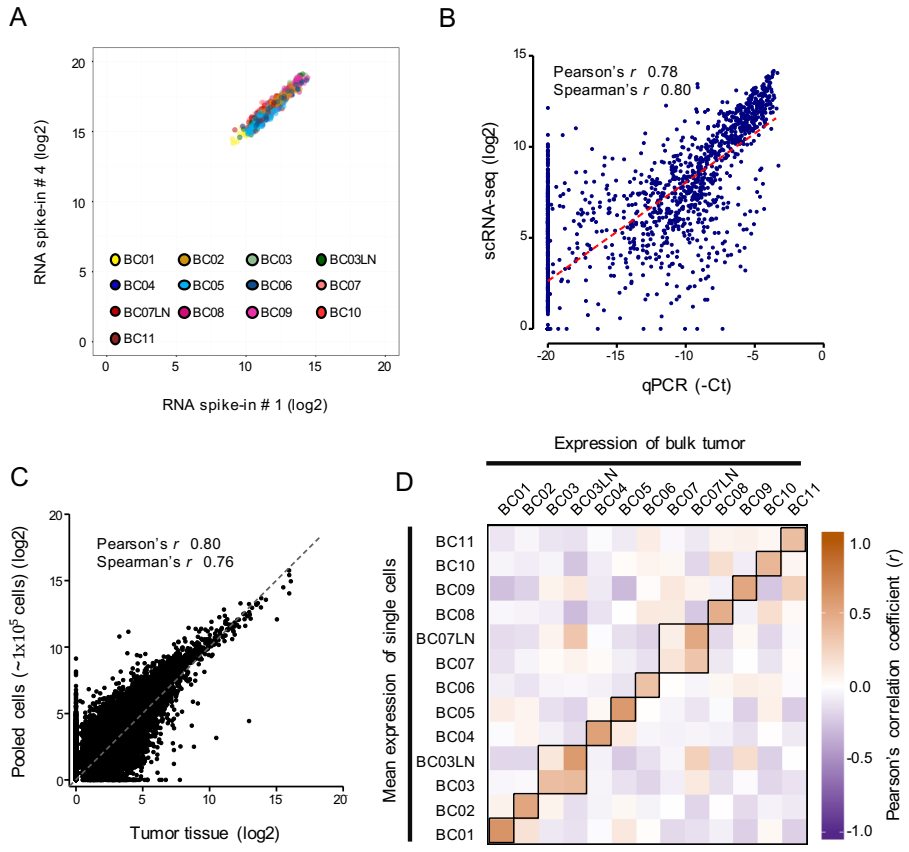

Figure 1-2. Reliability and representativeness of single-cell RNA-seq. (A) Consistent detection of normalized read counts for two array control RNA spike-ins in all single-cell samples. **(B)** Single-cell RNA-seq data showing a significant correlation with the matched qPCR results (Pearson's *r* 0.78). The linear regression result is drawn as a dashed line. **(C)** Significant correlation in the RNA-seq data between the tumor tissues and pooled tumor tissue isolates (Pearson's *r* 0.8). **(D)** Centered correlations between the average expressions of single cells and their matching bulk samples showing significant but partial representation of the bulk tumor by single cells (Pearson's *r* 0.16-0.63 with average 0.47, $p < 0.001$).

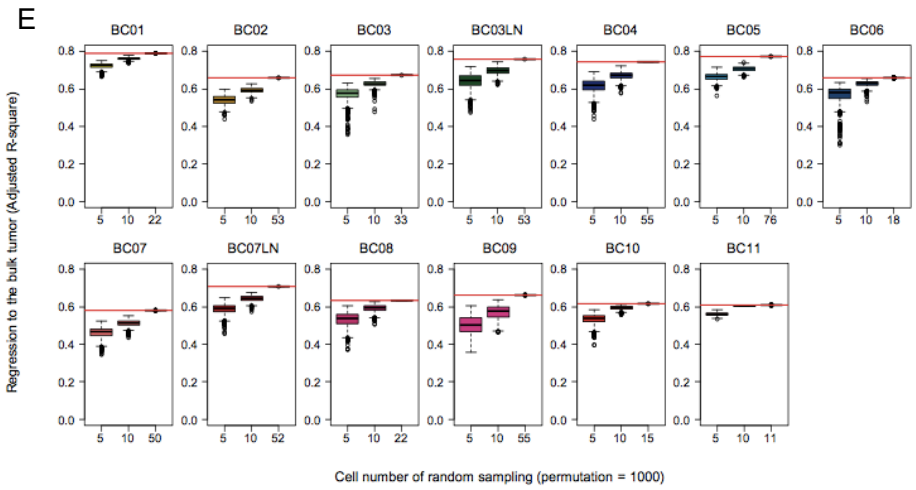

Figure 1-2. (E) Multiple regression analysis was performed using expression levels of each single cell as the explanatory variable to predict the expression level of bulk tumors. Adjusted R-squares of multiple regression analysis were calculated by random sampling of single cells with 1,000 iterations. Horizontal red lines represent maximum adjusted R-square values.

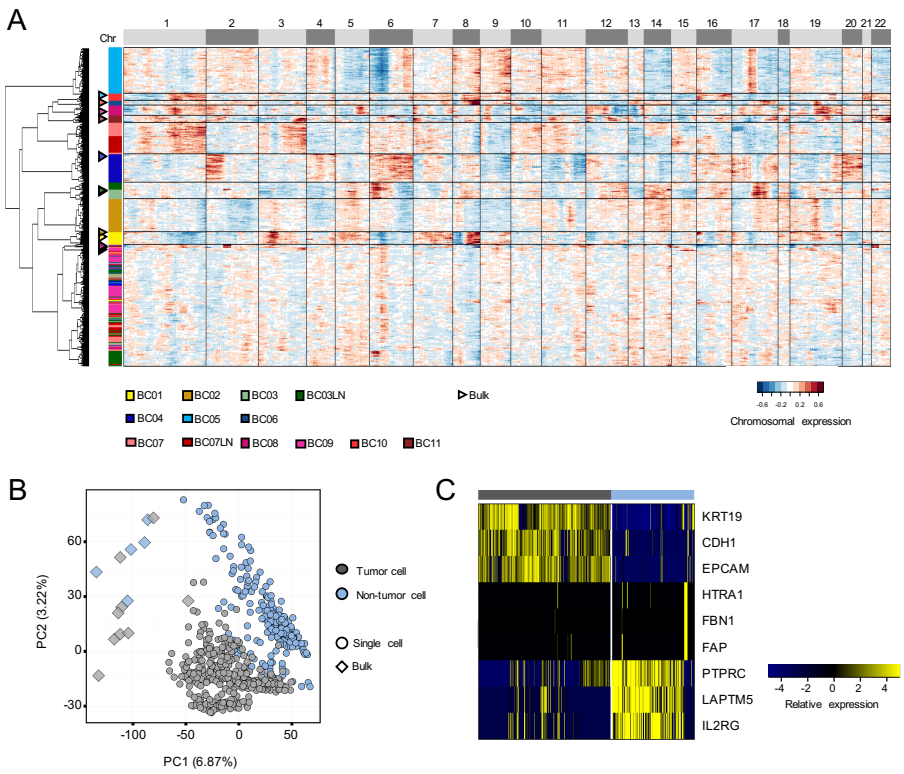

Figure 1-3. Chromosomal expression patterns separate tumor and non-tumor cells. (A) Unsupervised hierarchical clustering of the chromosomal gene expression pattern which reflects genomic copy number alterations mainly separates the tumor cell groups from the non-tumor cell cluster. Tumor cells have the patient-specific patterns whereas non-tumor cells are not. Each row represents profile of individual cells or matched bulk tumor tissues and triangle indicates bulk. Moving averages of 150 genes were calculated to infer the chromosomal expression pattern for each chromosome. **(B)** PCA shows the separation of tumor and non-tumor groups. **(C)** Representative gene expression levels (KRT19, CDH1, and EPCAM for epithelial cell; HTRA1, FBN1, and FAP for stromal cells; PTPRC, LAPTM5, and IL2RG for immune cells) confirmed the cell-type definition.

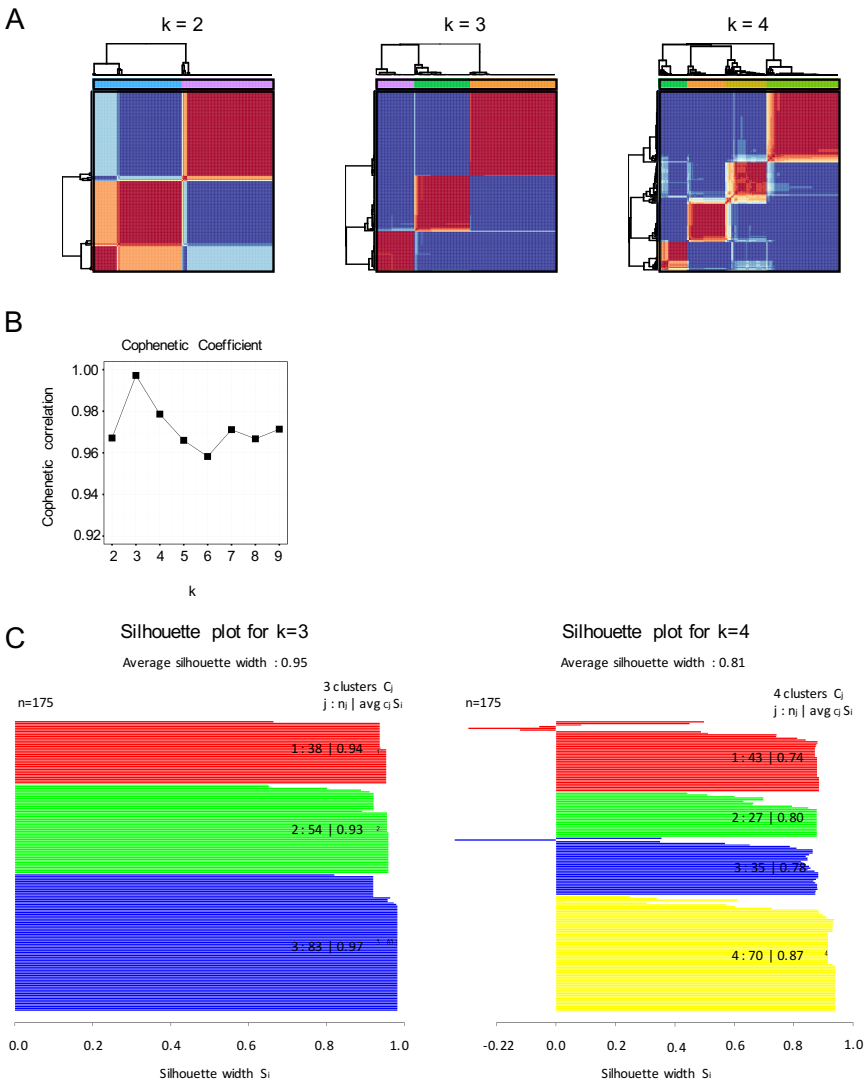

Figure 1-4. Subgrouping of non-tumor cells by NMF clustering. (A) NMF clustering analysis with immune gene sets for k = 2, k = 3, and k = 4. **(B)** The curve of cophenetic correlation coefficients. A peak was detected at k = 3, suggesting an optimal cluster of 3. **(C)** The silhouette width of each single cell for the k = 3 and k = 4 clusters. The k = 3 cluster had the highest average silhouette width (average S_i = 0.95) relative to the other clusters.

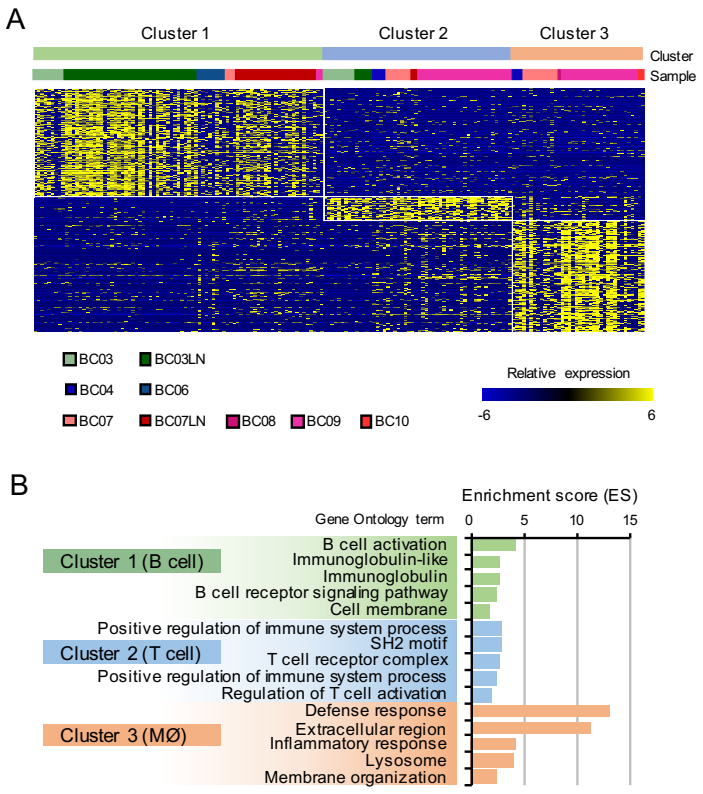

Figure 1-5. Identification of immune cell populations in breast tumor microenvironment. (A, B) Immune cell clusters were characterized by gene ontology terms. The cluster-specific genes extracted by likelihood-ratio test (LRT) **(A)** were associated with B cells, T cells, or macrophages (MØ) respectively **(B)**.

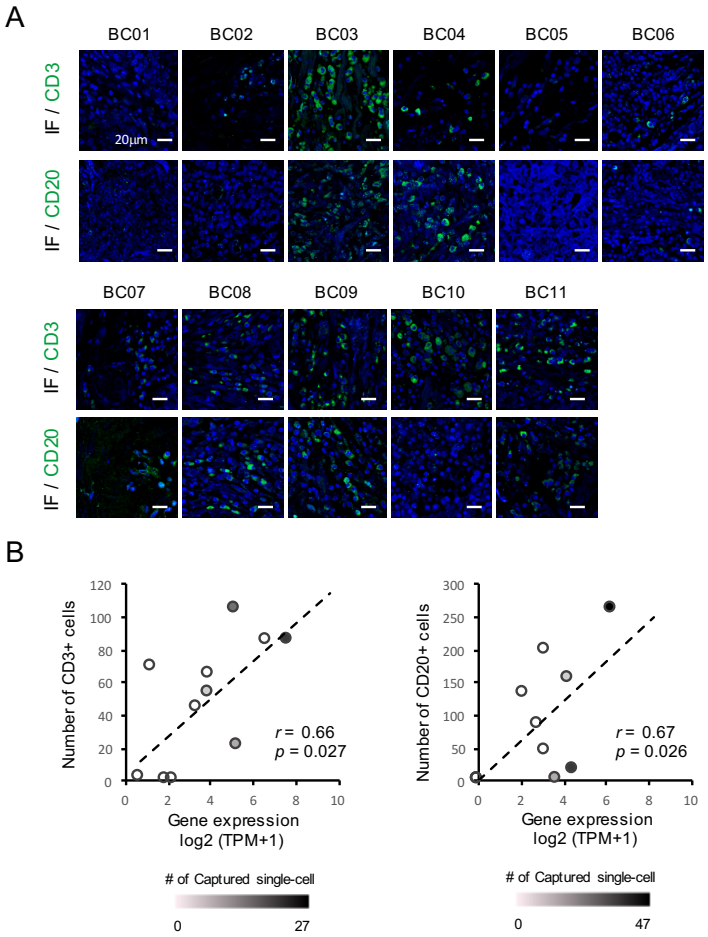

Figure 1-6. Immunostaining confirms the proportion of T or B lymphocytes in tumor tissues. (A) Immunofluorescence staining (IF) for CD3 or CD20 showing the infiltration of T or B lymphocytes in tumor tissues. Scale bar, 20μm. **(B)** Immunofluorescence staining results show significant correlations with gene expression in bulk tumor samples (Pearson's r, 0.66 for CD3 and 0.67 for CD20). The linear regression result is drawn as a dashed line. Number of captured single-cells is marked with a color key.

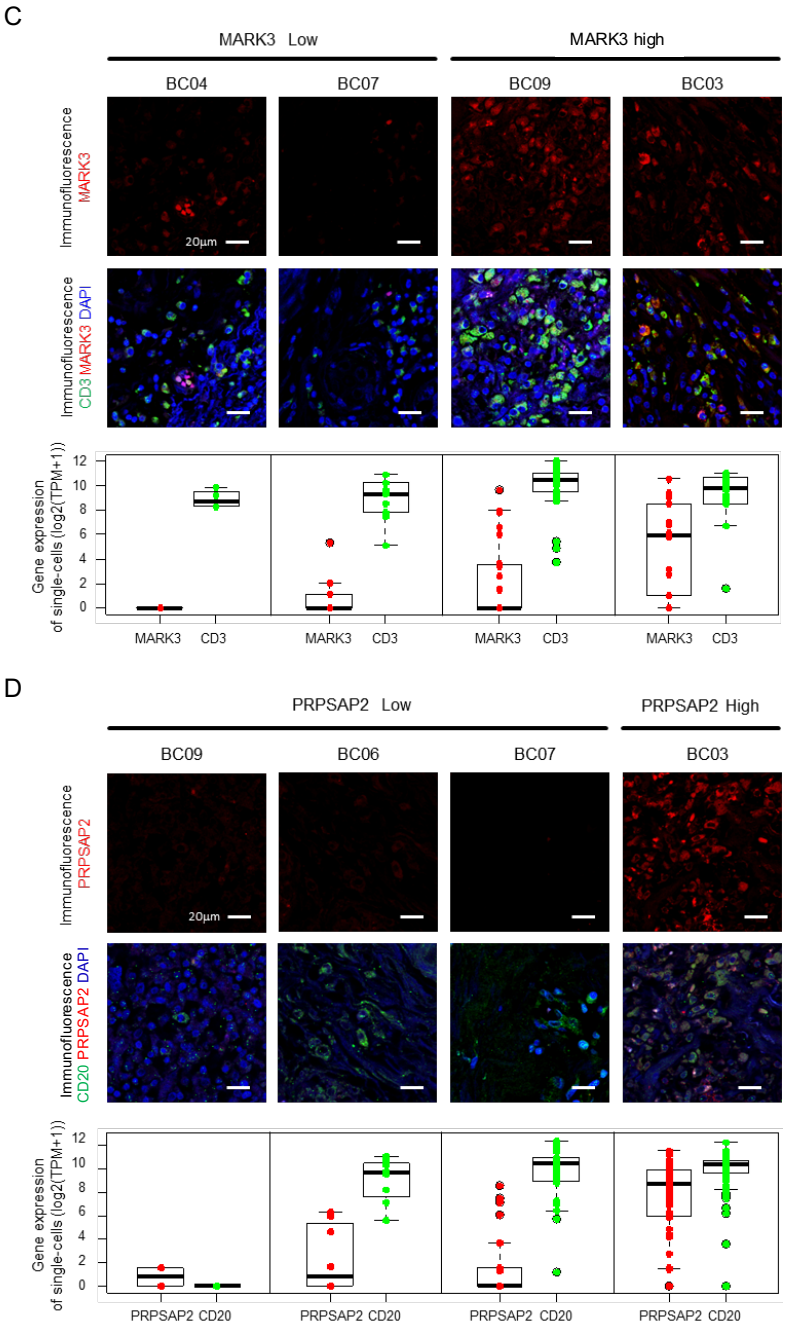

Figure 1-6. (C) MARK3 or **(D)** PRPSAP2 were co-stained with the T cell marker CD3 or B cell marker CD20 respectively. Single cell level gene expression is presented at the bottom for comparison. Scale bar, 20μm.

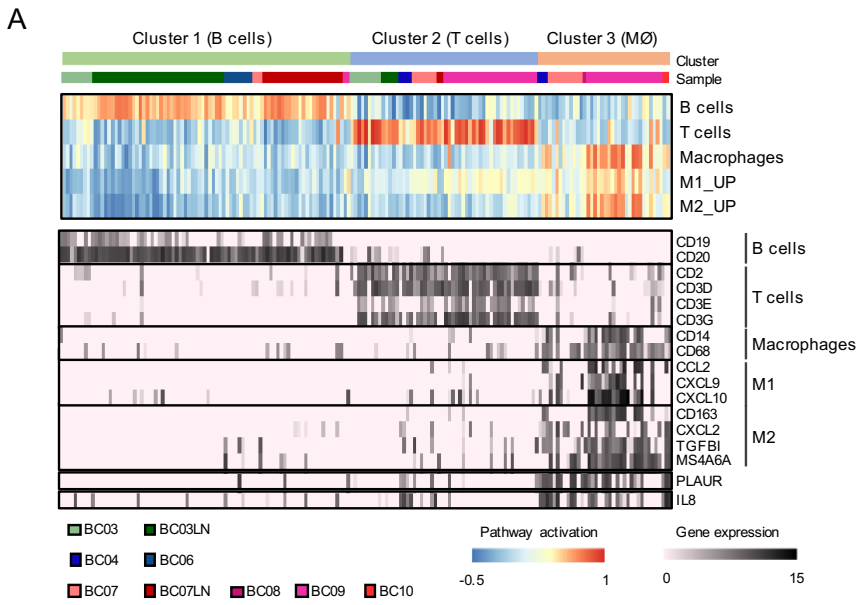

Figure 1-7. Heterogeneity within tumor-infiltrating immune cells. (A) Gene set variation analysis (GSVA) using immune gene sets (upper panel) shows characteristics of three clusters. The third cluster, expected as tumor associated macrophages expressed high levels of M2-type genes.

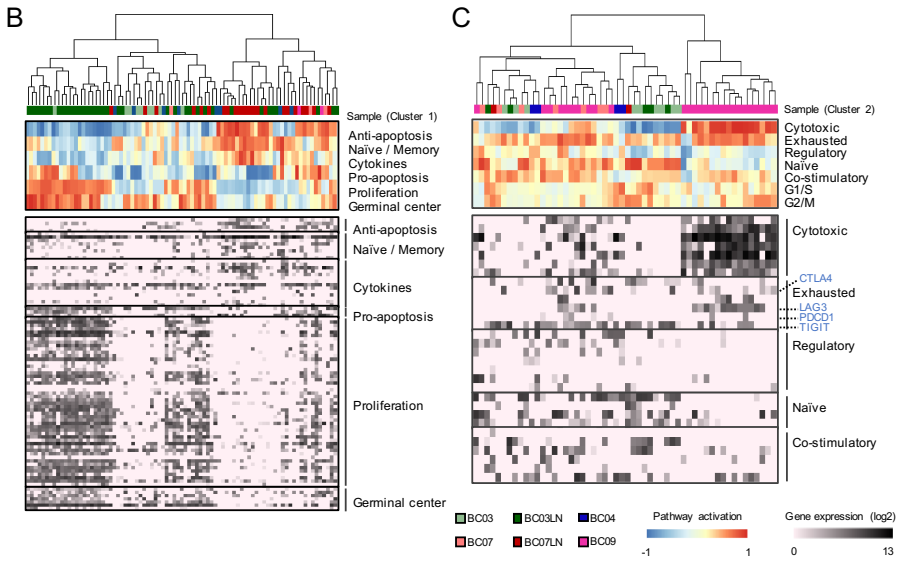

Figure 1-7. (B, C) Hierarchical clustering using GSVA enrichment scores (upper panel) for B lymphocytes **(B)** and T lymphocytes **(C)** was performed. Two subclasses of B lymphocytes were identified, one with an expression signature of germinal center B cells and the other with that of naïve B lymphocytes. Among the T lymphocytes, there are cytotoxic and exhausted cells derived from the BC09 sample and the other cells are non-cytotoxic naïve T cells. Gene expression profiles for the used gene sets are presented in the lower panel.

Table 1-1. Summary of breast cancer specimens.

Patient index	BC01 ER+	BC02 ER+	BC03 ER+/HER+	BC04 HER2+	BC05 HER2+	BC06 HER2+	BC07 TNBC	BC08 TNBC	BC09 TNBC	BC10 TNBC	BC11 TNBC
Age	66	72	72	67	46	67	71	67	53	82	47
Pathologic stage	pT1N0 (IA)	pT3N1 (IIIA)	pT2N1 (IIB)	pT2N0 (IIA)	ypT1N1mi (IB)	T2N1 (IIB)	pT1N3 (IIIC)	pT2N0 (IIA)	pT2N0 (IIA)	pT2N2 (IIIA)	T2N0 (IIA)
Molecular subtype	Luminal A	Luminal A	Luminal B	HER2-enriched	HER2-enriched	HER2-enriched	Basal-like	Basal-like	Basal-like	Basal-like	Basal-like
Immunohistochemistry											
ER	positive	positive	positive	negative	negative	negative (weak positive)	negative	negative	negative	negative	negative
PR	negative	positive	positive	negative	negative	negative (weak positive)	negative	negative	negative	negative	negative
HER2	2+/3	1+/3	3+/3	3+/3	3+/3	3+/3	1+/3	2+/3	1+/3	1+/3	1+/3
HER2 FISH	negative	-	-	-	-	-	-	negative	-	-	-
No. of single cells											
*Tumor	26 (22)	56 (53)	37 (33)	59 (55)	77 (76)	25 (18)	51 (50)	23 (22)	60 (55)	16 (15)	11 (11)
Lymph node	-	-	55 (53)	-	-	-	53 (52)	-	-	-	-
Lymph node metastasis	0/18	1/13	2/18	0/11	1/4	1/18	16/35	0/3	0/3	9/21	0/6

ER, estrogen receptor; IDC, invasive ductal carcinoma; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor; FISH, fluorescence in situ hybridization

* the number of total cells (the number of analysed cells)

Table 1-2. Information of primers used for quantitative PCR

RefSeq ID	Gene Symbol	Forward Primer	Reverse Primer
NM_002046	GAPDH	AGGTCGGAGTCAACGGATTT	TGACGGTGCCATGGAATTTG
NM_001101	ACTB	ACTCTTCCAGCCTTCCTTCC	CGTACAGGTCTTTGCGGATG
NM_000125	ESR1	GACAGGGAGCTGGTTCACA	GACCTGATCATGGAGGGTCAAA
NM_000926	PGR	AGCCAAGAAGAGTTCCTCTGTA	TTGACTTCGTAGCCCTTCCA
NM_004448	ERBB2	ACAGGGAAAGCTGTGGGAAA	TACGCCTCCAACACACTGAA
NM_005228	EGFR	AGGTGAAAACAGCTGCAAGG	CCAGAAGGTTGCACTTGTCC
NM_000633	BCL2	ATGTGTGTGGAGAGCGTCAA	GTGCCGGTTCAGGTACTCA
NM_002417	MKI67	AGAGTAACGCGGAGTGTCA	CTTGACACACACATTGTCCTCA
NM_000044	AR	CTTTGCAGCCTTGCTCTCTA	TCTGGTCGTCCACGTGTAA
NM_000546	TP53	GACTGTACCACCATCCACTACA	AAAGCTGTTCCGTCCCAGTA
NM_001012270	BIRC5	GGACCACCGCATCTCTACAT	GAAACACTGGGCCAAGTCTG
NM_018014	BCL11A	AACCCCAGCACTTAAGCAAAC	GGCCGTGGTCTGGTTCA
NM_000224	KRT18	TCCCATGTCCCAGTCAATTCC	TACCTGGGAGGGGATGTTCA
NM_002276	KRT19	GGAGGTGTCATTGGAGCTGAA	AGCAGCTTCCACCACTTCAA
NM_002051	GATA3	CACGGTGCAGAGGTACCC	AGGGTAGGGATCCATGAAGCA
NM_004360	CDH1	AGTGCCAACTGGACCATTCA	TCTAAGGCCATCTTTGGCTTCA
NM_001067	TOP2A	GGTGTGGAACTAGAAGGCCT	TCTGTTTCTCGTGGAGGGAC
NM_000201	ICAM1	AACCCCACAGTCACCTATGG	TTCTGAGACCTCTGGCTTCG
NM_001432	EREG	TTGTTTGCATGGACAGTGCA	GCTTAAAGGTTGGTGGACGG
NM_080921	PTPRC	GTGGCTTAAACTCTTGGCATTT	GGGAAGGTGTTGGGCTTT
NM_002354	EPCAM	CGTCAATGCCAGTGTACTTCA	TTCTGCCTTCATCACCAAACA
NM_201442.2	C1S	TACGGGGTTTGCTGCATACTA	AGTGGCTACAAGGGACATCTAC
NM_000064.3	C3	GGCGTGTTCGTGCTGAATAA	CCGATGTCTGCCTTCTCCA
NM_000346	SOX9	GTGCTCAAAGGCTACGACTG	AGAAGTCTCCAGAGCTTGCC

Table 1-3. List of immune gene sets used for NMF clustering

Gene set name	B cells	T cells	T helper cells	CD8 T cells	Cytotoxic cells	NK cells	Dendritic cells	Eosinophils	Macrophages	Mast cells	Neutrophils
Genes	MS4A1	PRKCQ	ICOS	CD8B	KLRD1	LOC643313	CD209	IL5RA	MARCO	PRG2	CSF3R
	TCL1A	CD3D	LRBA	CD8A	KLRF1	GAGE2	CCL17	KCNH2	CXCL5	CTSG	CYP4F3
	HLA-DOB	CD3G	ITM2A	PF4	GNLY	ZNF747	HSD11B1	TKTL1	SCG5	TPSAB1	VNN3
	PNOC	CD28	FAM111A	PRR5	CTSW	XCL1	CCL13	EMR1	SULT1C2	SLC18A2	FPRL1
	KIAA0125	LCK	PHF10	SF1	KLRB1	XCL2	CCL22	CCR3	MSR1	MS4A2	KCNJ15
	CD19	TRAT1	NUP107	LIME1	KLRK1	AF107846	PPFIBP2	ACACB	CTSK	CPA3	MME
	CR2	BCL11B	SEC24C	DNAJB1	NKG7	SLC30A5	NPR1	THBS1	PTGDS	TPSB2	IL8RA
	IGHG1	CD2	NAP1L4	ARHGAP8	GZMH	SGMS1	CD1B	GALC	COLEC12	TPSB2	IL8RB
	FCRL2	TRBC1	BATF	GZMM	SIGIRR	MCM3AP	VASH1	RNU2	GPC4	GATA2	FCGR3B
	BLK	TRAC	ASF1A	SLC16A7	ZBTB16	TBXA2R	F13A1	CLC	PCOLCE2	HDC	DYSF
	COCH	ITM2A	FRYL	SFRS7	RUNX3	CDC5L	CD1E	HIST1H1C	CHIT1	LOH11CR2A	FCAR
	OSBPL10	SH2D1A	FUSIP1	APBA2	APOL3	LOC730096	MMP12	CYSLTR2	KAL1	SIGLEC6	CEACAM3
	IGHA1	CD6	YME1L1	C4orf15	RORA	FUT5	FABP4	HRH4	CLEC5A	ELA2	HIST1H2BC
	TNFRSF17	CD96	TRA	LEPROTL1	APBA2	FGF18	CLEC10A	RNASE2	ME1	CMA1	HPSE
	ABCB4	NCALD	RPA1	ZFP36L2	WHDC1L1	MRC2	SYT17	CAT	DNASE2B	PGDS	CPPED1
	BLNK	GIMAP5	UBE2L3	GADD45A	DUSP2	RP5-886K2.1	MS4A6A	LRP5L	CCL7	MLPH	CREB5
	GLDC	TRA	ANP32B	MYST3	GZMA	SPN	CTNS	SYNJ1	FN1	ADCYAP1	S100A12
	MEF2C	CD3E	DDX50	ZEB1		PSMD4	GUCA1A	THBS4	CD163	SLC24A3	TNFRSF10C
	IGHM	SKAP1	C13orf34	ZNF609		PRX	CARD9	GPR44	GM2A	CALB2	SLC22A4
	FAM30A		PPP2R5C	C12orf47		FZR1	ABCG2	KBTBD11	SCARB2	KIT	TECPR2
	SPIB		SLC25A12	THUMPD1		ZNF205	CD1A	HES1	BCAT1	TAL1	SLC25A37
	BCL11A		ATF2	VAMP2		AL080130	PPARG	ABHD2	RAI14	ABCC4	BST1
	GNG7		CD28	ZNF91		ZNF528	RAP1GAP	TIPARP	COL8A2	PPM1H	CRISPLD2
	IGKC		GOLGA8A	ZNF22		MAPRE3	SLC7A8	SMPD3	APOE	MAOB	G0S2
	CD72		IFNG	TMC6		BCL2	GSTT1	MYO15B	CHI3L1	HPGD	SIGLEC5
	MICAL3		LTA	FLT3LG		NM_017616	PDXK	TGIF1	ATG7	SCG2	CD93
	BACH2		APBB2	CDKN2AIP		ARL6IP2	FZD2	RRP12	CD84	PTGS1	MGAM
	IGL		DOK5	TSC22D3		PDLIM4	CSF1R	IGSF2	FDX1	CEACAM8	ALPL
	CCR9		IL12RB2	TBCC		TRPV6	HS3ST2	RCOR3	MS4A4A	MPO	FPR1
	QRSL1		APOD	RBM3		LDB3	CH25H	EPN2	SGMS1	NR0B1	PDE4B
	DTNB		ZBTB32	ABT1		ADARB1	LMAN2L	C9orf156	EMP1	LOC339524	LILRB2
	HLA-DQA1		CD38	C19orf6		SMEK1	SLC26A6	SIAH1	CYBB		
	SCN3A		CSF2	CAMLG		TCTN2	BLVRB		CD68		
	SLC15A2		CTLA4	PPP1R2		TINAGL1	NUDT9				
			CD70	AES		IGFBP5	PREP				
			DPP4	KLF9		ALDH1B1	TM7SF4				
			EGFL6	PRF1		NCR1	TACSTD2				
			BST2			KIR3DL2	CD1C				
			DUSP5			SPON2	CCL1				
			LRP8			KIR2DL3	EBI3				
			IL22			GZMB	INDO				
			DGKI			KIR3DS1	LAMP3				
			CCL4			KIR3DL1	OAS3				
			GGT1			TCC38	IL3RA				
			LRRN3			PMEPA1					
			SYNGR3			IL21R					
			ATP9A			KIR3DL3					
			BTG3			KIR2DS5					
			CMAH			KIR2DS2					
			HBEGF			GTF3C1					
			SGCB			KIR2DS1					
			PMCH			S1PR5					
			AHI1			DUSP4					
			PTGIS			RRAD					
			CXCR6			PLA2G6					
			EVI5			NIBP					
			IL26			FOXJ1					
			MB			MARCH6					
			NEIL3			MADD					
			GSTA4			LPCAT4					
			PHEX			MPPED1					
			SMAD2			MUC3B					
			CENPF								
			ANK1								
			ADCY1								
			LOC728210								
			LAIR2								
			SNRPD1								
			MICAL2								
			DHFR								
			WDHD1								
			BIRC5								
			SLC39A14								
			HELLS								
			LIMA1								
			CDC25C								
			CDC7								
			GATA3								

Table 1-4. Cluster-specific gene lists identified at a single cell resolution

A. B cell upregulated

Gene symbol	Fold change	P-value (LRT test)	AUC (ROC test)	Gene symbol	Fold change	P-value (LRT test)	AUC (ROC test)	Gene symbol	Fold change	P-value (LRT test)	AUC (ROC test)
IGLC1	12.37	3.9.E-12	0.70	DCAF12	4.43	1.6.E-11	0.71	SEL1L3	2.69	4.9.E-11	0.79
FCRLA	8.11	0.0.E+00	0.84	ABCA6	4.26	1.8.E-11	0.72	BCAR3	2.67	4.9.E-10	0.73
RN7SL627P	7.98	3.7.E-15	0.74	GPR18	4.16	8.9.E-16	0.81	SSBP2	2.66	3.3.E-12	0.75
RN7SL639P	7.98	3.7.E-15	0.74	PAX5	4.11	0.0.E+00	0.89	BCAS4	2.66	4.2.E-15	0.77
GCSAM	7.81	7.8.E-16	0.75	CTA-250D10.23	4.09	0.0.E+00	0.77	FAM3C	2.65	2.0.E-08	0.72
VNN2	7.78	0.0.E+00	0.77	MCTP2	4.06	1.6.E-12	0.71	TMED8	2.65	2.7.E-11	0.75
AICDA	7.77	0.0.E+00	0.78	CDK14	4.04	7.7.E-12	0.74	FAM210A	2.64	3.0.E-08	0.70
IGHG4	7.68	0.0.E+00	0.80	BRI3BP	4.01	2.5.E-11	0.72	KIAA0922	2.64	2.1.E-12	0.80
VPREB3	7.17	0.0.E+00	0.84	LPP	3.93	5.1.E-14	0.76	PTK2	2.60	3.4.E-08	0.73
RGS13	7.16	0.0.E+00	0.85	NCF1	3.84	1.7.E-12	0.71	MZB1	2.60	5.5.E-08	0.71
SNX29P1	7.05	0.0.E+00	0.80	ENTPD4	3.83	6.1.E-10	0.72	KBTBD8	2.55	1.8.E-13	0.78
E2F5	6.77	1.8.E-13	0.70	DTX1	3.81	7.9.E-13	0.72	FAM208B	2.49	1.7.E-08	0.73
PNOC	6.68	7.2.E-14	0.72	POU2AF1	3.75	0.0.E+00	0.82	LAT2	2.46	4.0.E-11	0.79
FCRL1	6.65	2.4.E-14	0.72	RRM2B	3.70	2.1.E-10	0.72	BIK	2.40	9.7.E-14	0.77
AC079767.4	6.55	2.2.E-12	0.70	RAB30	3.70	0.0.E+00	0.87	HLA-DOB	2.37	6.9.E-09	0.72
ELL3	6.48	0.0.E+00	0.78	FAM3C2	3.51	1.7.E-11	0.70	ADAM28	2.35	7.0.E-10	0.76
IGHG1	6.37	0.0.E+00	0.86	RRAS2	3.51	4.3.E-14	0.79	SEC14L1	2.34	1.4.E-07	0.75
EBF1	6.11	0.0.E+00	0.82	SHCBP1	3.50	4.0.E-10	0.72	PHF6	2.33	3.3.E-08	0.73
CCDC144B	6.09	0.0.E+00	0.80	CTD-2369P2.2	3.44	1.7.E-14	0.82	MEF2C	2.33	0.0.E+00	0.88
IGKC	6.09	5.6.E-16	0.85	DNAJC10	3.39	1.6.E-12	0.77	UBE2J1	2.31	4.0.E-09	0.76
CD79A	6.02	0.0.E+00	0.92	PRKD3	3.35	2.5.E-14	0.82	CLIC4	2.29	3.0.E-06	0.71
CD19	5.89	0.0.E+00	0.78	HAUS8	3.23	1.0.E-07	0.71	TRAK1	2.25	7.9.E-07	0.70
LINC00877	5.87	2.1.E-13	0.71	KLHL6	3.22	2.6.E-10	0.75	RPRD1B	2.22	2.0.E-08	0.72
CCDC144A	5.86	4.4.E-16	0.76	LRMP	3.10	0.0.E+00	0.86	SWAP70	2.12	0.0.E+00	0.87
STAG3	5.85	1.6.E-11	0.71	HDAC9	3.10	7.8.E-16	0.82	LY9	2.12	1.9.E-10	0.75
RALGPS2	5.80	3.3.E-16	0.76	P2RX5	3.09	6.3.E-12	0.72	FAM76B	2.08	2.3.E-09	0.74
ZNF608	5.64	0.0.E+00	0.78	COBLL1	3.06	7.5.E-15	0.79	PIK3C2B	2.07	1.0.E-10	0.71
CD22	5.57	0.0.E+00	0.89	ST6GAL1	3.04	0.0.E+00	0.84	RHOH	2.04	2.1.E-12	0.81
PRPSAP2	5.40	5.6.E-16	0.80	DGKD	2.97	6.1.E-09	0.72	CCNB1	2.04	1.0.E-06	0.72
MS4A1	5.39	0.0.E+00	0.97	BLNK	2.91	6.4.E-12	0.79	MRPS27	2.03	2.6.E-08	0.72
P2RY8	5.33	1.1.E-13	0.71	BANK1	2.81	5.2.E-13	0.74	SPATS2	2.03	2.5.E-07	0.71
IGHG3	5.11	0.0.E+00	0.86	SMIM14	2.79	0.0.E+00	0.81	SNX22	2.02	2.8.E-12	0.72
GMDS	5.06	2.3.E-11	0.71	EIF2AK3	2.75	2.7.E-13	0.79	TMEM156	2.02	4.9.E-07	0.72
NEIL1	4.96	1.1.E-12	0.74	EAF2	2.74	0.0.E+00	0.85	IGLC3	2.01	1.9.E-08	0.71
SNX29P2	4.93	0.0.E+00	0.91	LYPLAL1	2.73	1.6.E-07	0.71	NR4A1	2.00	6.6.E-06	0.71
PLCG2	4.85	0.0.E+00	0.83	TEX9	2.72	3.2.E-12	0.71				
CD79B	4.55	0.0.E+00	0.89	FANCA	2.72	2.3.E-09	0.71				

B. T cell upregulated

Gene symbol	Fold change	P-value (LRT test)	AUC (ROC test)	Gene symbol	Fold change	P-value (LRT test)	AUC (ROC test)	Gene symbol	Fold change	P-value (LRT test)	AUC (ROC test)
IFNG	6.19	7.7.E-10	0.71	RORA	3.63	3.7.E-10	0.75	CD96	2.78	6.0.E-14	0.81
CD3G	5.06	0.0.E+00	0.90	SH2D1A	3.43	1.1.E-16	0.79	MAF	2.77	1.0.E-07	0.73
CD3D	4.46	0.0.E+00	0.94	PRKCH	3.30	1.0.E-09	0.75	CCL5	2.75	3.8.E-06	0.72
INPP4B	4.07	3.0.E-11	0.72	TRAT1	3.24	5.0.E-10	0.71	TIGIT	2.73	1.1.E-09	0.75
CD2	4.06	0.0.E+00	0.95	FYN	3.07	2.7.E-13	0.82	SLA	2.68	6.8.E-09	0.74
ITK	3.90	8.9.E-10	0.73	ARAP2	2.93	5.3.E-09	0.75	TRAC	2.66	2.2.E-16	0.87
STAT4	3.77	9.6.E-12	0.77	ITM2A	2.82	1.1.E-16	0.82	TRBC2	2.35	0.0.E+00	0.91
IL32	3.64	0.0.E+00	0.91	CD3E	2.81	2.2.E-15	0.78	RARRES3	2.01	3.2.E-06	0.73

Table 1-4. (continued)

C. Macrophage upregulated

Gene symbol	Fold change	P-value (LRT test)	AUC (ROC test)	Gene symbol	Fold change	P-value (LRT test)	AUC (ROC test)	Gene symbol	Fold change	P-value (LRT test)	AUC (ROC test)
C1QB	9.79	2.3.E-12	0.72	PLXDC2	3.73	4.6.E-11	0.74	CST3	2.64	0.0.E+00	0.93
CD163	9.28	1.1.E-15	0.73	CREG1	3.70	3.8.E-11	0.75	VAMP3	2.62	9.5.E-09	0.74
IL1B	9.14	2.8.E-14	0.72	MNDA	3.69	8.1.E-13	0.81	VAMP5	2.61	1.6.E-06	0.74
FCGR3A	8.61	4.5.E-14	0.75	KIAA1598	3.68	2.1.E-08	0.73	BEST1	2.60	6.4.E-09	0.79
CCL2	8.52	1.9.E-11	0.72	GPNMB	3.67	7.3.E-09	0.72	RBM47	2.58	2.1.E-07	0.76
TMEM176B	8.24	0.0.E+00	0.80	MYOF	3.66	9.3.E-10	0.77	CEBPB	2.52	2.0.E-07	0.70
CXCL9	7.97	7.3.E-12	0.73	TIMP1	3.65	2.0.E-07	0.75	PSAP	2.51	1.7.E-09	0.82
FCGR1A	7.69	1.4.E-11	0.71	AXL	3.65	3.6.E-15	0.80	TGFBI	2.50	6.9.E-14	0.83
APOC1	7.59	5.2.E-15	0.79	LAIR1	3.60	1.4.E-10	0.75	CXCL16	2.49	1.0.E-08	0.73
FN1	7.51	3.3.E-16	0.83	CD68	3.60	1.6.E-14	0.85	TIMP2	2.48	2.3.E-12	0.78
TMEM176A	7.09	3.9.E-12	0.72	MS4A6A	3.60	1.1.E-16	0.84	ITM2B	2.47	2.2.E-08	0.80
S100A9	6.72	2.1.E-10	0.73	TYROBP	3.55	0.0.E+00	0.88	CPVL	2.44	2.2.E-07	0.70
APOE	6.68	1.4.E-12	0.71	RASSF4	3.54	3.0.E-11	0.77	IGSF6	2.44	9.8.E-12	0.75
SLAMF8	6.66	2.9.E-11	0.71	DSE	3.51	3.3.E-16	0.78	HEXB	2.42	1.5.E-09	0.79
CD14	6.35	4.2.E-13	0.76	FTL	3.50	0.0.E+00	0.88	CTSC	2.41	8.6.E-10	0.81
CXCL10	6.05	5.4.E-10	0.75	KCTD12	3.23	1.2.E-13	0.77	BRI3	2.39	1.9.E-07	0.73
FCGR2A	5.72	7.2.E-12	0.75	FCER1G	3.21	0.0.E+00	0.87	GPX1	2.36	7.1.E-11	0.84
RP11-1143G9.4	5.63	1.3.E-15	0.81	FCGRT	3.21	5.3.E-11	0.79	MARCKS	2.36	1.6.E-07	0.77
HNMT	5.55	4.4.E-12	0.71	CTSB	3.17	3.3.E-13	0.85	CAPG	2.33	7.8.E-08	0.79
PLBD1	5.48	7.3.E-14	0.70	HLA-DQB2	3.09	5.4.E-05	0.72	IFITM3	2.33	1.1.E-10	0.83
A2M	5.35	0.0.E+00	0.82	SLC8A1	3.05	2.8.E-07	0.70	S100A11	2.31	1.8.E-11	0.88
CXCL2	5.06	1.8.E-10	0.74	IER3	3.04	4.2.E-08	0.72	HLA-DQA2	2.29	1.3.E-07	0.77
MAFB	5.02	1.1.E-16	0.80	MFSD1	3.03	2.7.E-11	0.82	LST1	2.28	6.4.E-10	0.80
MSR1	4.87	1.8.E-12	0.75	PLAUR	3.00	5.4.E-13	0.81	TNFSF13B	2.25	4.4.E-11	0.80
PLAU	4.81	1.6.E-09	0.72	IL18	3.00	2.8.E-10	0.74	RNASET2	2.23	1.7.E-07	0.77
CTSL	4.69	3.1.E-14	0.82	ANKRD22	2.96	1.5.E-09	0.72	FTLP3	2.23	2.4.E-10	0.80
SIRPA	4.66	1.9.E-13	0.76	ATP6AP1	2.93	9.2.E-07	0.71	LPCAT2	2.20	2.6.E-11	0.71
FGL2	4.62	6.2.E-12	0.76	SOD2	2.90	9.7.E-08	0.77	CD63	2.20	1.8.E-07	0.78
LYZ	4.55	2.3.E-15	0.82	GLUL	2.86	8.5.E-09	0.79	CCL4	2.20	2.7.E-04	0.70
SERPING1	4.51	0.0.E+00	0.88	IFIT1	2.83	4.4.E-06	0.71	LILRB4	2.19	1.5.E-09	0.75
SCARB2	4.40	4.2.E-11	0.76	IFI27	2.82	1.5.E-13	0.87	HLA-DRB6	2.16	6.2.E-08	0.72
CLEC7A	4.36	7.9.E-15	0.84	CCND1	2.79	7.4.E-08	0.71	MS4A7	2.13	2.2.E-06	0.72
AIF1	4.20	6.1.E-14	0.83	IFIT3	2.71	4.5.E-05	0.71	ATOX1	2.12	3.0.E-06	0.75
PILRA	4.19	6.1.E-11	0.71	GNAQ	2.71	4.1.E-11	0.77	DAB2	2.09	2.2.E-15	0.81
C15orf48	4.00	4.0.E-12	0.76	IL8	2.71	8.8.E-08	0.76	NAGK	2.02	5.2.E-07	0.76
FAM26F	3.97	3.0.E-09	0.71	NPC2	2.70	1.0.E-15	0.90	IFI30	2.02	6.6.E-09	0.81
SERPINA1	3.80	9.7.E-11	0.75	NRP1	2.70	1.7.E-08	0.73				
RIN2	3.77	5.6.E-10	0.73	RNASE6	2.66	4.3.E-07	0.73				

Table 1-5. Gene lists used in gene set variation analysis

A. M1/M2 gene sets

Gene set name	M1_UP	M2_UP
Genes	CCR7	GPR86
	IL2RA	P2RY5
	IL15RA	TGFBR2
	IL7R	HRH1
	CXCL11	TLR5
	CCL19	DCL-1
	CXCL10	MSR1
	CXCL9	CXCR4
	TNF	DECTIN1
	CCL5	P2RY14
	CCL15	DCSIGN
	IL12B	CLECSF13
	IL15	MS4A6A
	TRAIL	CD36
	IL6	MS4A4A
	CCL20	MRC1
	PBEF1	IGF1
	ECGF1	CCL23
	BCL2A1	CCL18
	FAS	CCL13
	BIRC3	SLC21A9
	GADD45G	SLC4A7
	HSXIAPAF1	SLC38A6
	SLC7A5	CTSC
	SLC21A15	HEXB
	SLC2A6	LIPA
	SLC31A2	ADK
	INDO	HNMT
	PLA1A	TPST2
	OASL	CERK
	CHI3L2	HS3ST2
	HSD11B1	LTA4H
	AK3	CA2
	SPHK1	ALOX15
	PFKFB3	HS3ST1
	PSME2	TGFBI
	PFKP	SEPP1
	PSMB9	CHN2
	PSMA2	FN1
	OAS2	FGL2
	PTX3	GAS7
	CSPG2	EGR2
	APOL3	MAF
	IGFBP4	
	APOL1	
	PDGFA	
	EDN1	
	APOL2	
	INHBA	
	APOL6	
	HESX1	
	IRF1	
	ATF3	
	IRF7	

B. B cell signature gene sets

Gene set name	Anti-apoptosis	Naïve / Memory	Cytokines	Pro-apoptosis	Proliferation	Germinal center
Genes	BCL2	BMI1	CCL5	BIK	ASK	BCL7A
	CASP8	CD24	CCR1	FAS	AURKA	CD27
	FAIM3	CD69	CCR6	LGALS1	AURKB	CD36
	TNFSF10	CR1	CCR7		AURKC	CD80
		ENTPD1	CLR1		BUB1B	HGF
		FCER2	GPR9		CCNA2	MME
		FCGR2B	IFNGR1		CCNB1	MYBL1
		FCGRT	IL10RB		CCNB2	RGS13
		IGHD	IL15		CCNE2	TNFSF8
			IL24		CCNF	
			IL2RB		CDC20	
			IL2RG		CDC25B	
			IL4R		CDC45	
			IL6		CDC6	
			IL8		CDK1	
			TGFB1		CDK5	
			TGFB3		CDKN2C	
			TGFBR2		CENPA	
			TNFRSF1B		CENPE	
			TNFSF11		CENPF	
			XCL1		CHEK1	
					CIP2	
					DEEPEST	
					E2F5	
					ECA39	
					FOXA1	
					GADD45A	
					GADD45B	
					GTSE1	
					HEC1	
					KIF11	
					KIF22	
					KIF23	
					MAD2L1	
					MCM3	
					MCM4	
					MCM6	
					MCM7	
					MKI67	
					NCAPD2	
					NDN	
					NEK2	
					PCNA	
					PLK4	
					PTTG1	
					RABGAP1	
					RAD17	
					RFC3	
					RGL2	
					RPA3	
					SIL-TAL1	
					TFDP1	
					TMPO	
					TPX2	
					TTK	
					UBE2C	
					WEE1	
					ZW10	

Table 1-5. (continued)

C. T cell signature gene sets

Gene set name	Cytotoxic	Exhausted	Regulatory	Naïve	Costimulatory	G1/S	G2/M
Genes	CST7	BTLA	IL2RA	CCR7	ICOS	ATAD2	ANLN
	GZMA	CTLA4	IL4R	LEF1	CD226	BLM	ANP32E
	GZMB	HAVCR2	IL7	SELL	SLAMF1	BRIP1	AURKA
	IFNG	LAG3	TGFB1	TCF7	TNFRSF14	CASP8AP2	AURKB
	NKG7	PDCD1	TGFB3		TNFRSF25	CCNE2	BIRC5
	PRF1	TIGIT	TGFBI		TNFRSF9	CDC45	BUB1
	TNFSF10		TGFBR1			CDC6	CBX5
						CDCA7	CCNB2
						CHAF1B	CDC20
						CLSPN	CDC25C
						DSCC1	CDCA2
						DTL	CDCA3
						E2F8	CDCA8
						EXO1	CDK1
						FEN1	CENPA
						GINS2	CENPE
						GMNN	CENPF
						HELLS	CKAP2
						MCM2	CKAP2L
						MCM4	CKAP5
						MCM5	CKS1B
						MCM6	CKS2
						MLF1IP	CTCF
						MSH2	DLGAP5
						NASP	ECT2
						PCNA	FAM64A
						POLA1	G2E3
						POLD3	GAS2L3
						PRIM1	GTSE1
						RAD51	HJURP
						RAD51AP1	HJURP
						RFC2	HMGB2
						RPA2	HMMR
						RRM1	HN1
						RRM2	KIF11
						SLBP	KIF20B
						TIPIN	KIF23
						TYMS	KIF2C
						UBR7	LBR
						UHRF1	MKI67
						UNG	NCAPD2
						USP1	NDC80
						WDR76	NEK2
							NUF2
							NUSAP1
							PSRC1
							RANGAP1
							SMC4
							TACC3
							TMPO
							TOP2A
							TPX2
							TTK
							TUBB4B
							UBE2C

RESUTLS (PART-II)

Uncovering heterogeneous polarization levels of tumor-associated macrophages using single-cell RNA sequencing

1-1. Separation of macrophages in advanced gastric cancer environment by single-cell RNA-seq

Single cell transcriptome data from the peritoneal metastatic ascites of four gastric cancer patients and the cerebrospinal fluid metastasis of the AGC04 patient (AGC04CSF) were acquired without any prior selection (Table 2-1). Because all of the patients diagnosed as being in an advanced stage of cancer with ascites or additional regional metastasis, all samples were collected after receiving treatment according to pathological marker expression.

To identify the macrophage cells in our AGC data, the CNA pattern was initially inferred from each cell in order to separate tumor cells (Figure 2-1A). Hierarchical clustering of inferred chromosomal expressions displayed four tumor groups (except AGC04) and one non-tumor group. Unsupervised clustering using an shared nearest neighbor (SNN) algorithm classified clusters of each cell and PCA analysis reconfirmed the separation of the clusters by PC1 (Figure 2-1B and C). Moreover, a non-linear dimensional reduction method called t-Distributed Stochastic Neighbor Embedding (tSNE) (50), which is frequently used to visualize clear clustering over PCA, perfectly mirrored the groups in Figure 2-1A (Figure 2-1B and C).

In general, the cells contained in malignant ascites could be tumor cells, tumor-associated immune cells, peritoneum-derived mesothelial cells originated from primary tumor or peritoneal metastasis (29-31). The cell types of the two clusters were identified by the expression of the several marker genes, including EPCAM for epithelial cells, CA125 for peritoneal

mesothelial cells, and CD68 for macrophages (Figure 2-1D). Most of the cells in Cluster 1 strongly expressed EPCAM, whereas most of the cells in Cluster 2 expressed CD68, suggesting that Cluster 2 is a cluster of macrophages. Notably, the non-tumor mesothelial cells that expressed CA125 were clustered in a tumor cluster without EPCAM expression. These few mesothelial cells were distinctly grouped by chromosomal expressions. Altogether, only the macrophage cells in the tumor microenvironment were separated by *in silico* analysis.

2-2. Construction of M1 and M2 single-cell transcriptome profiles

The M2-type macrophages are well known to have anti-inflammatory functions and an increased M2/M1 ratio is associated with a poor prognosis in ovarian or lung cancer (14,15). To evaluate the obtained single-cell TAM type, reference transcriptomes for M1- or M2-type macrophages were constructed by single-cell RNA-seq from two healthy donors (Figure 2-2). A total of 97 individual M1 cells (38 from donor 1 and 59 from donor 2) and 45 individual M2 cells (18 from donor 1 and 27 from donor 2) were sequenced. The differentiation and polarization of PBMC monocytes as M1 or M2 was validated using cell morphology and marker expression (Figure 2-2A and B).

Unsupervised PCA showed that the macrophage cells are mainly separated by macrophage type and then separated by origin donor (Figure 2-2C). Marker gene expressions of single-cell data reaffirmed the cell type and found

that marker gene expression patterns and levels were very similar between the two donors despite the different number of cells (Figure 2-2D). As a result, M1 and M2 transcriptomes were correctly produced, but partial differences between the samples according to the donor still remained.

2-3. M2 scoring using reference transcriptomes

In a previous report, the gene expression profiles of various states of macrophages were already established by microarray-based data (41). The authors also provided the differentially expressed genes (DEGs) between the M1 and M2 populations (Table 1-5A). To examine the DEGs of M1 and M2 with a single-cell resolution, an LRT test was performed first based on zero-inflated data, and gene lists were extracted from each donor (Figure 2-3A). Considering the differences derived from the donors, the 148 and 38 genes repeatedly extracted from the two donors were selected as M1 cell-specific and M2 cell-specific signature genes. In the bulk profiling of DEGs, only 20 out of 54 genes in M1 and nine out of 43 genes in M2 were detected as DEGs in single-cell profiles. The gene expression levels of the 97 bulk-derived DEGs were examined in single cells, and the 29 overlapping genes showed higher expression changes with significance (Student's t-test $P < 1e5$) (Figure 2-3B), suggesting that these 29 persistently observed genes may provide new markers for both in bulk and single-cell levels.

Next, the level of M2-polarization in TAMs was estimated through GSVA using our reference transcriptomes to obtain DEG sets. When evaluating M2-polarization by using TAMs, both AGC TAM and BC TAM seemed to have a

balanced polarization at a similar level (Figure 2-4A). After applying reference transcriptomes to set the extreme point of M1 or M2, however, both AGC TAM and BC TAM showed M2-biased intermediate polarization (Figure 2-4A and B). Furthermore, the same analysis was repeatedly performed using different set of genes. When using single-cell derived gene sets, even though both AGC and BC TAMs showed an intermediate level of M2-polarization, strong M2 polarization in AGC TAMs was observed. Taken together, this analysis performed with reference profiles may assist in the prevention of a distorted perspective produced by internal comparison, and the usage of single cell-derived gene sets provides more objective results in single-cell data.

2-4. Heterogeneous polarization levels of TAMs in various tumor types

Many studies have proven the high correlation between M2-like TAM and poor survival rate (10-15), but a negative result has also been reported in some tumor types such as colorectal cancer (CRC) (51). Therefore, the levels of M2-like polarization of TAMs in three tumor types — AGC, BC, and CRC — were compared using collected single-cell data (1,32) (Figure 2-5). All of the TAMs from these three tumor types presented an intermediate level of M2-like polarization with variable median scores (Figure 2-5A). In comparison, most M2-like macrophages were AGC TAMs. Using the single-cell derived

gene sets shortened the range of wide-spreading M1 or M2 reference cells, which also supports the requirement for appropriate gene sets.

Clustering the cells by the dimensional reduction method predominantly divided them by their origin tissue (Figure 2-5A). In order to define the state of TAMs with continuous M2-like scores, cells were sorted using a DDRTree algorithm (42,43) to order cells using gene expression profiles (Figure 2-5B). Cell ordering using core DEGs classified a total of three states: State 1 with reference M2 cells and some TAMs, State 2 with M1 cells from donor 1 and others with TAMs, and State 3 with only the M1 cells of donor 2. As shown in the previous results, all AGC TAMs were classified as State 1, which is the M2 group. In case of BC and CRC, however, the M1-like polarized TAMs were also observed in State 2. All of our results support the mixed existence of M1-like and M2-like TAMs in the microenvironment. Additionally, the results showed the various proportions of M2-like TAMs and the heterogeneity in the polarization level according to the tumor type.

Table 2-1. Clinical information of AGC patients for single-cell sequencing.

Patient index	AGC01	AGC02	AGC03	AGC04	AGC04CSF
Metastasis	Peritoneal ascites	Peritoneal ascites	Peritoneal ascites	Peritoneal ascites	Cerebrospinal fluid
Prognostic markers					
FGFR2 amp	negative	negative	positive	positive	-
MET high	negative	positive	negative	negative	-
***No. of single cells**	25 (25)	48 (48)	34 (34)	39 (22)	34 (33)

* the number of total cells (the number of analysed cells)

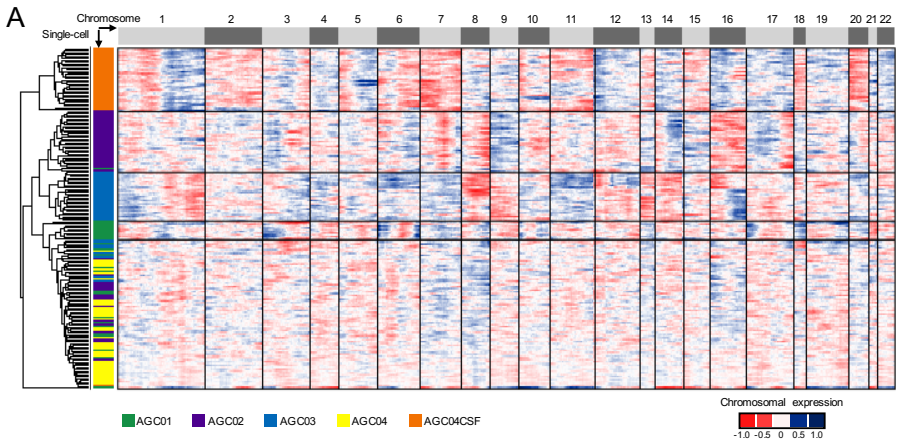

Figure 2-1. Separation of tumor-associated macrophages from malignant cells. (A) Non-cancerous cells are distinguished from cancerous cells by using the CNA patterns estimated from the chromosomal expression levels. For each chromosome, the chromosomal gene expression pattern was estimated using the moving average of 150 genes. The level of chromosomal expression in each cell was normalized with an average value of GTEx normal tissue. Each row represents single cells.

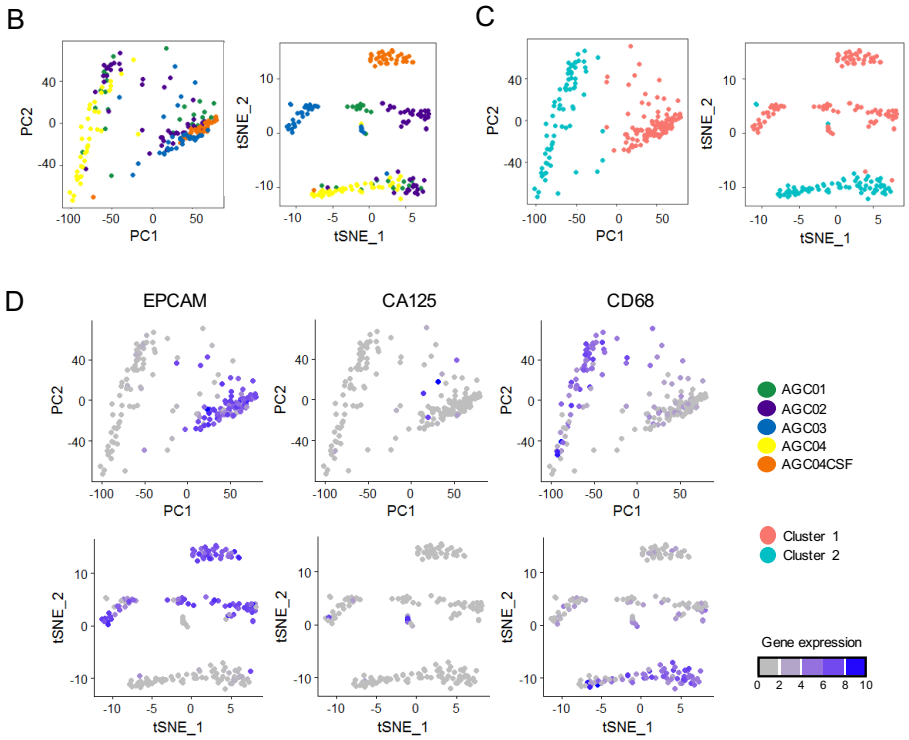

Figure 2-1. (B) PCA (upper) shows that the most important component by which cells are clustered is their cancerous property. Further dimension reduction using tSNE (lower) separates non-cancerous cell clusters and patient-specific carcinoma cell clusters. **(C)** The classification of the cluster of each cell by a shared nearest neighbor (SNN) algorithm. **(D)** The expression of marker genes for epithelial cells (EPCAM), peritoneal mesothelial cells (CA125), and macrophages (CD68) shows clusters of carcinoma cells and macrophages.

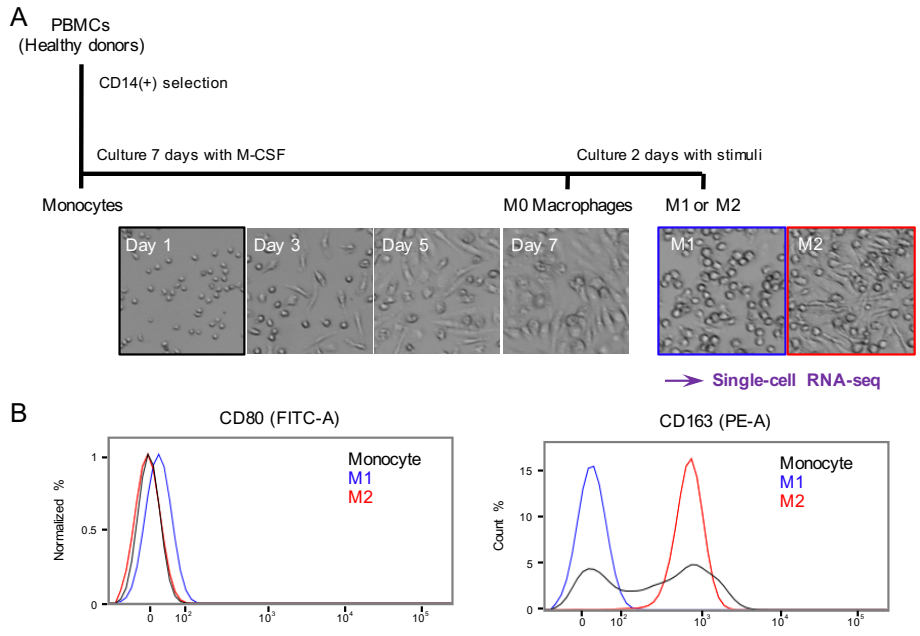

Figure 2-2. Construction of the single-cell transcriptomes of M1 and M2 macrophages derived from normal PBMCs. (A) The experimental design and morphology of polarized M1 and M2. **(B)** The expression of macrophage markers were examined by FACS. CD80, an M1-specific marker, is only expressed in M1-polarized macrophages, whereas the M2-specific marker CD163 is more highly expressed in M2-polarized macrophages than in M1. The examined samples are colored with borders (black for the monocyte, blue for the M1 macrophage, and red for the M2 macrophage) in (A).

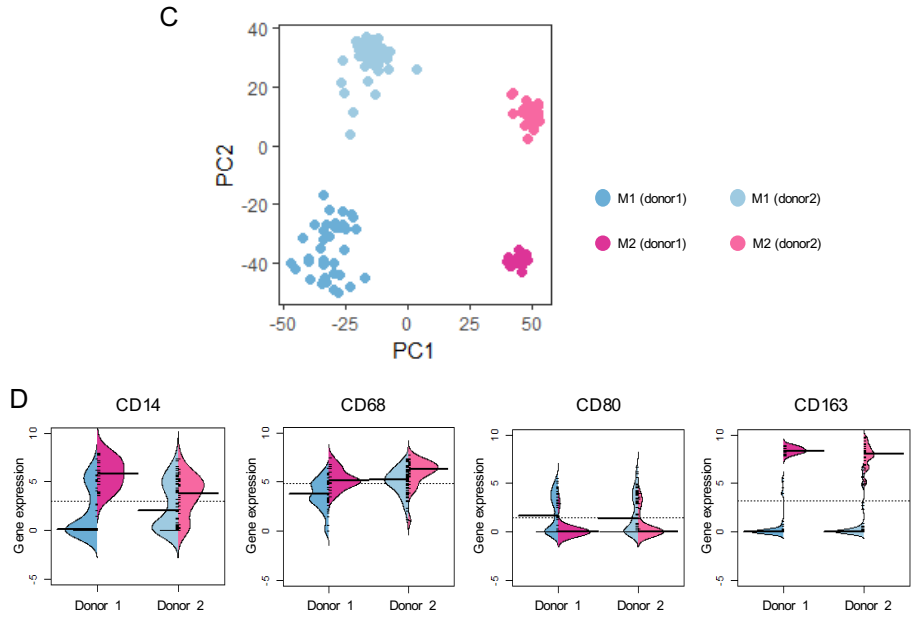

Figure 2-2. (C) Unsupervised PCA primarily separates M1 and M2 macrophage cells by PC1. **(D)** The gene expression level of macrophage markers in single-cells. The expression level is calculated using log2(TPM/10+1), and the thick black line indicates the median value.

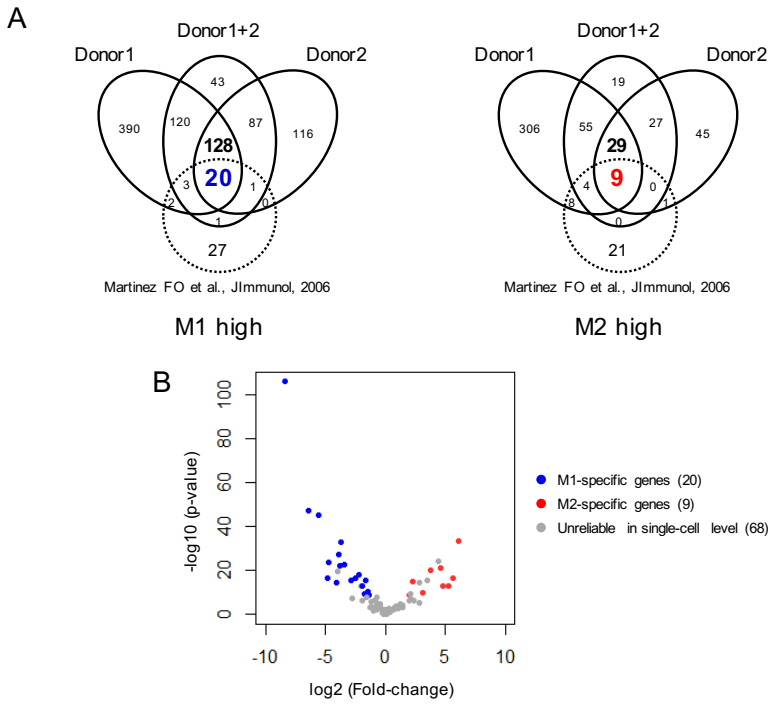

Figure 2-3. Extracting differentially expressed genes from M1 and M2 single-cell transcriptomes. (A) The overlapping differentially expressed genes (DEGs) from both donors leave 186 core DEGs for M1-specific or M2-specific signatures. When compared with the published gene set derived from M1 and M2 bulks (Martinez FO et al.), only 29 genes are repeatedly extracted in single-cell and in bulk. **(B)** The validated 29 genes of the published gene set show a significant difference in expression between M1 and M2 single-cells (p < 1e5). The x-axis indicates a log2-fold-change in the average expression of M1 and M2 single-cells and the y-axis is a negative log10 of the p-value obtained by a Student's t-test.

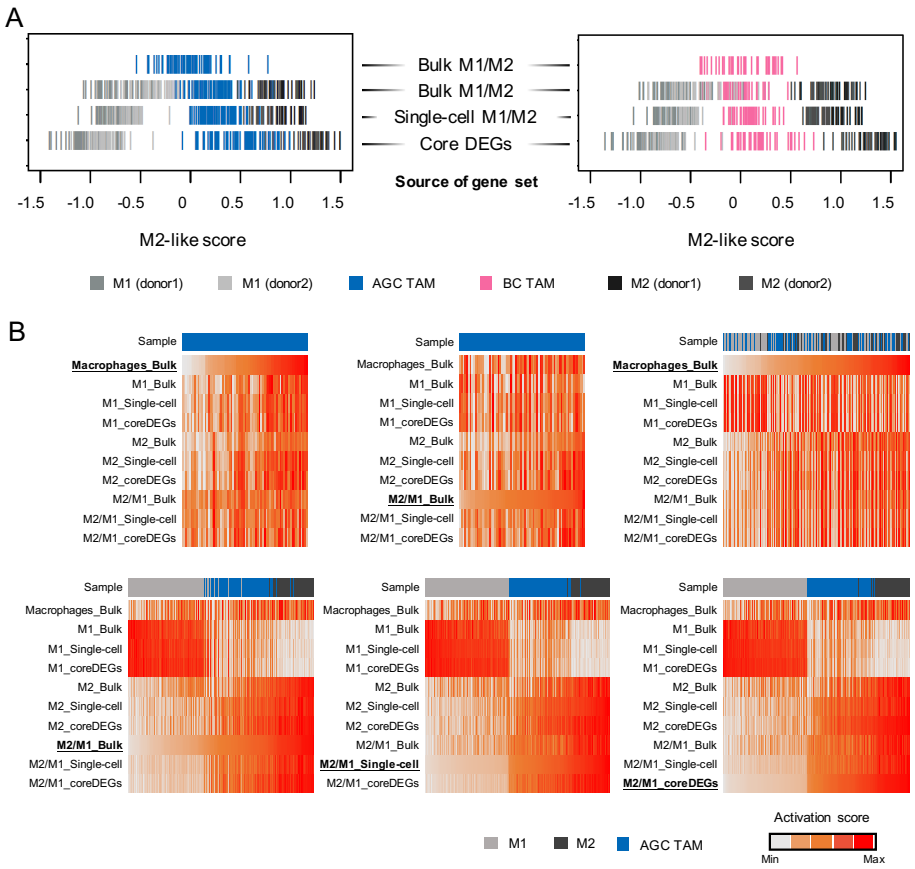

Figure 2-4. Application of the reference transcriptome gives more objective insight about the M2-like polarization level in TAMs. (A) Geneset variation analysis (GSVA) was performed in the TAMs both without (first rows) and with (2, 3, and 4 rows) the reference cells. The tendency of M2-like polarization is observed after adding the single-cell M1 and M2 cells as the reference. The M2-like score (x-axis) is calculated as the score of M2 DEGs minus the score of M1 DEGs. **(B)** Ordering cells by the diverse gene sets' GSVA score confirms that the newly constructed DEG lists derived from single-cells provide a more distinctive score for TAMs.

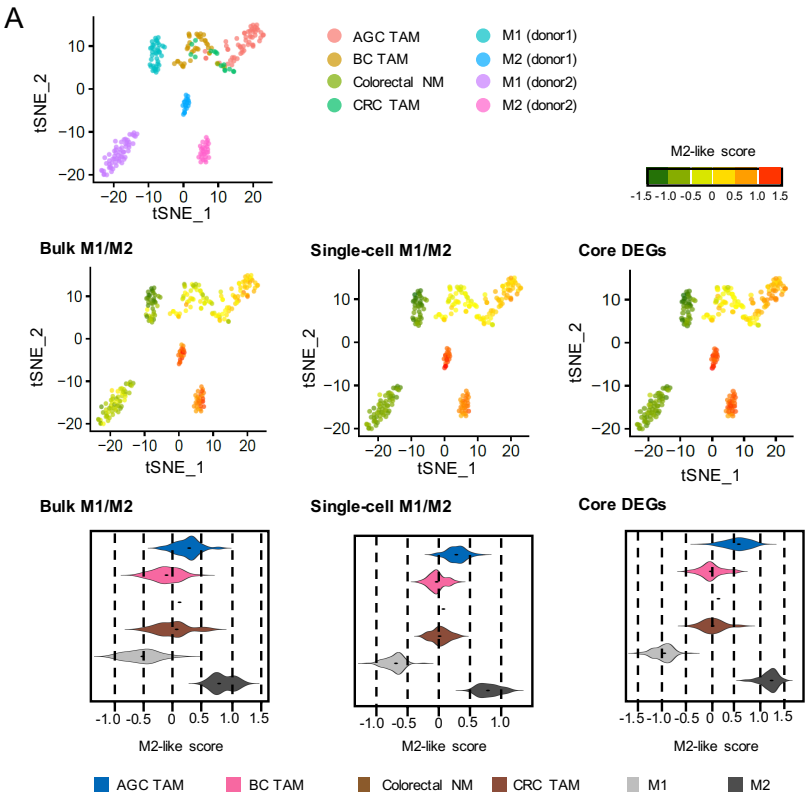

Figure 2-5. Heterogeneous M2-biased levels in the macrophages of various tumor types. (A) The unsupervised clustering by t-SNE (top) and the projection of M2-like scores (middle) of various TAMs together. The biased levels of macrophages by tumor origin are summarized in the violin plots (bottom). The median is marked with a black dot in the violin plot. The M2-like score (x-axis) is calculated as the score of M2 DEGs minus the score of M1 DEGs.

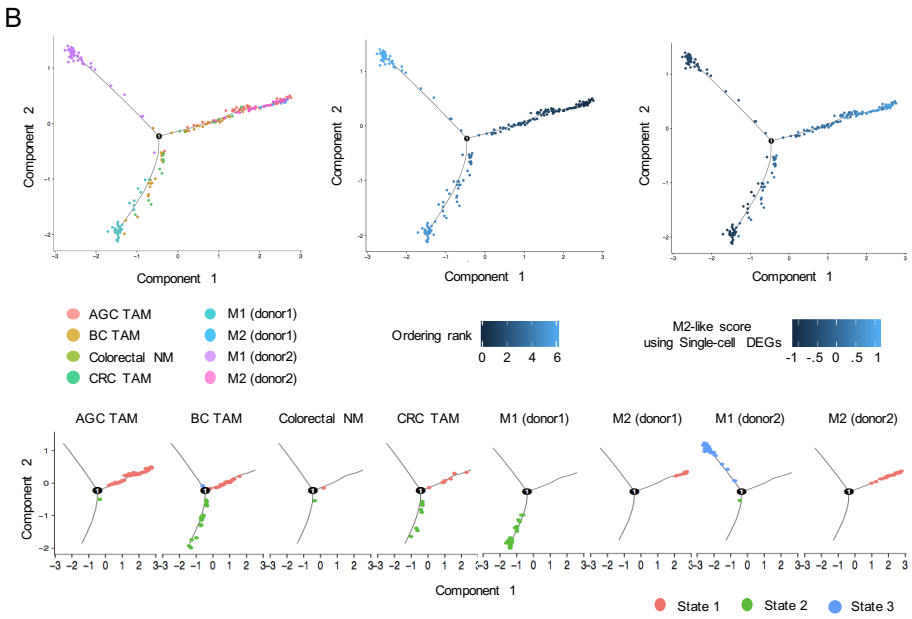

Figure 2-5. (B) The ordering of cells with M1/M2 single-cell DEGs using the DDRTree algorithm. The dispersion of macrophages into M1 or M2 is plotted at the bottom by tumor origin. The TAMs from AGC are predominantly polarized to the M2 type, whereas half of the TAMs from BC and CRC are more M1-like cells. Each cell is colored by its tumor origin (top left), pseudo time (top middle), M2-like score (top right), or state (bottom).

DISCUSSION

Conventionally, the study of tumor-infiltrating immune cells has been conducted using pre-defined immune populations based on known surface marker expression. Here, I used an inferred genomic feature, i.e. chromosomal expression pattern to separate tumor and immune cells without prior marker selection. Single-cell analysis without marker selection allows to simultaneous capture of various components in the tumor microenvironment. I could clearly separate the signatures of tumor and tumor-infiltrating immune cells using transcriptomic analysis of single-cell isolates, which together defined the characteristics of cancer. In the non-tumor cell analysis of breast cancer, I classified non-tumor cells into three immune cell types with activating and suppressive gene expression signatures. The reference single-cell transcriptomes of normal blood-derived M1 and M2 macrophages were utilized to present a more objective polarization level of TAM separated *in silico* from AGC. When comparing TAMs from various cancer types with reference transcriptomes, each tumor type showed a difference in M1/M2 balance. All of these results suggest dynamic immune cell interactions and a distinct immune system status in each tumor.

The role of the immune system in tumor progression has been extensively studied, and the results of these studies have provided the basis for successful immunotherapy in multiple cancer types (9,52). Tumors are thought to evade natural immune surveillance either by immunologic ignorance or by active suppression. T cell infiltration may be a key determinant by which the evasion

pathway operates. Among immune cell infiltrates mostly obtained from TNBC tumors, I found a number of T cells with high cytokine and chemokine expression, which is indicative of ongoing immune responses. Interestingly, these T cells also manifested immuno-suppressive phenotypes of exhausted (8,46) or regulatory T cells (47). Therapeutic strategies to overcome T cell exhaustion, i.e., immune checkpoint blockade, have been developed that target CTLA-4 and PD-1/PD-L1, and these strategies have demonstrated significant efficacy in treating melanoma and non-small cell lung cancer (9). Clinical trials of these treatment strategies for other cancer types, including breast cancer, are in progress as a mono- or combination therapy. Search for new immune checkpoint targets is ongoing and early clinical trials for additional checkpoint molecules such as LAG3 (53) have begun. In our single-cell dataset, most T cells expressed LAG3 and/or TIGIT (54) (Figure 1-7C), suggesting that they are potential targets for checkpoint inhibition.

Transforming anti-inflammatory M2 cells into pro-inflammatory M1 cells is one strategy for cancer treatment, and the re-polarization of M2 macrophages to the M1-type is possible by changing the microenvironment (55). In gastric cancer, there is a report showing that M1/M2 conversion by changing the *in vitro* culture condition is similar to the progression of the tumor's condition (56), suggesting the inverted possibility of M2 to M1 conversion. By comparing differentiated M1 and M2 single-cell transcriptomes, M1-specific and M2-specific genes were extracted, and applying these genes allowed for the objective estimation of the M2-polarization level and macrophage status. Macrophages collected from fluid

metastasis of AGC showed the most M2-biased features compared to other cancer types (Figure 2-5), but the reason for this remains uncertain. One possibility is that there exists a sampling bias caused by using samples harvested at the most malignant state, stage IV. Alternatively, these results may be derived from the difference in tumor types or tissue-specific properties. Most peritoneal macrophages are M2-type. The origin cell type could be identified by comparing ambiguous cells with normal cell transcriptomes of each cell type. In spite of some limitations, our results suggest that single-cell profiling could be used to discover new targets for M2/M1 conversion and to evaluate the state of TAM and any changes that occur.

Though there was no marker selection, I collected only a small number of tentative cancer-associated fibroblasts or epithelial cells, and no endothelial cells in breast cancer, probably due to the limitations of the cell isolation and capture methods. Similarly, few mesothelial cells and no TIL population were detected in advanced gastric cancer. The partial representation of bulk tumor transcriptomes by those of single cells (Figure 1-2D and E) also suggests limitations of sampling in the current approach. To overcome this limitation and to profile the entire tumor microenvironment, cell isolation techniques enabling large scale, unbiased sampling especially no size limitation need to be explored (39, 57).

Overall, considering the diverse immune cell types, single-cell expression profiling is particularly important for the accurate characterization of tumor-infiltrating immune cells and may provide some clues for immunotherapy.

Furthermore, the results of this study also demonstrate the requirement for large-scaled single-cell gene expression profiling with reduced sampling biases for the inclusive characterization of heterogeneous immune cells.

REFERENCES

1. Chung W, Eum HH, Lee HO, et al. Single-cell RNA-seq enables comprehensive tumor and immune cell profiling in primary breast cancer. Nat Commun 8, 15081 (2017)

2. Finak G, et al. Stromal gene expression predicts clinical outcome in breast cancer. Nat Med 14, 518-527 (2008).

3. Salgado R, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO 26, 259-271 (2015).

4. Galon J, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science 313, 1960-1964 (2006).

5. Luo Y, et al. Targeting tumor-associated macrophages as a novel strategy against breast cancer. J Clin Invest 116, 2132-2141 (2006).

6. Schwartz M, Zhang Y, Rosenblatt JD. B cell regulation of the anti-tumor response and role in carcinogenesis. Journal for immunotherapy of cancer 4, 40 (2016).

7. Mahmoud SM, Lee AH, Paish EC, Macmillan RD, Ellis IO, Green AR. The prognostic significance of B lymphocytes in invasive carcinoma of the breast. Breast cancer research and treatment 132,

545-553 (2012).

8. Wherry EJ et al. Molecular signature of CD8+ T cell exhaustion during chronic viral infection. Immunity 27, 670-684 (2007).

9. Postow MA, Callahan MK, Wolchok JD. Immune Checkpoint Blockade in Cancer Therapy. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 33, 1974-1982 (2015).

10. Noy R, Pollard JW. Tumor-associated macrophages: from mechanisms to therapy. Immunity 41, 49-61 (2014).

11. Wolf H, et al. The immune contexture in cancer prognosis and treatment. Nat Rev Clin Oncol 7, 399-416 (2017).

12. KJ Kim, et al. Prognostic Implication of M2 Macrophages Are Determined by the Proportional Balance of Tumor Associated Macrophages and Tumor Infiltrating Lymphocytes in Microsatellite-Unstable Gastric Carcinoma. PLoS One 10, e0144192 (2015).

13. Oris CM, et al. Macrophages within NSCLC tumour islets are predominantly of a cytotoxic M1 phenotype associated with extended survival. Eur Respir J 33, 118-126 (2009).

14. Zhang, et al. A high M1/M2 ratio of tumor-associated macrophages is associated with extended survival in ovarian cancer patients. Journal of Ovarian Research 7, 19 (2014)

15. Zhang J, et al. High Infiltration of Tumor-Associated Macrophages Influences Poor Prognosis in Human Gastric Cancer Patients, Associates With the Phenomenon of EMT. Medicine (Baltimore) 95, e2636 (2016).

16. Navin NE. Tumour evolution inferred by single-cell sequencing. Nature 7, 90-94 (2011).

17. Patel AP, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396-1401 (2014).

18. Tirosh I, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189-196 (2016).

19. Kim KT, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome biology 16, 127 (2015).

20. Navin NE. Delineating cancer evolution with single-cell sequencing. Sci Transl Med 7, 296fs229 (2015).

21. Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ. Clinical application of the 70-gene profile: the MINDACT trial. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 26, 729-735 (2008).

22. Sparano JA, Paik S. Development of the 21-gene assay and its

application in clinical practice and clinical trials. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 26, 721-728 (2008).

23. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature 490, 61-70 (2012).

24. Roshandel G, et al. Endoscopic screening for esophageal squamous cell carcinoma. Arch Iran Med 16, 358-365 (2013).

25. Simone Mocellin, Daunia Verdi, Karen A Pooley, Donato Nitti, et al. Genetic variation and gastric cancer risk: a field synopsis and meta-analysis. Gut 8, 1209-1219 (2015).

26. Mairi H. McLean & Emad M. El-Omar. Genetics of gastric cancer. Nat Rev Gastroenterol Hepatol 11, 664-674 (2014).

27. Changwon Kang, Ji-Joon Song, Jaeok Lee, Mi Young Kim, et al. Epigenetics: An emerging player in gastric cancer. World J Gastroenterol 20, 6433-6447 (2014).

28. DU-GUAN FU. Epigenetic alterations in gastric cancer (Review). Mol Med Rep 12, 3223-3230 (2015).

29. Emma Kipps, David S. P. Tan, Stan B. Kaye, et al. Meeting the challenge of ascites in ovarian cancer: new avenues for therapy and research. Nat Rev Cancer 13, 273-282 (2013).

30. Bertrum Sheid. Angiogenic effects of macrophages isolated from

ascitic fluid aspirated from women with advanced ovarian cancer. Cancer Lett 62, 153-158 (1992).

31. J Li, et al. The impact of inflammatory cells in malignant ascites on small intestinal ICCs' morphology and function. J Cell Mol Med 19, 2118-2127 (2015).

32. Li H, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nature Genetics 49, 708–718 (2017).

33. Haibe-Kains B, et al. A three-gene model to robustly identify breast cancer molecular subtypes. J Natl Cancer Inst 104, 311-325 (2012).

34. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21 (2013).

35. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).

36. DeLuca DS, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 28, 1530-1532 (2012).

37. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC bioinformatics 11, 367 (2010).

38. Bindea G, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity 39,

782-795 (2013).

39. Macosko EZ, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202-1214 (2015).

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

41. Martinez FO, Gordon S, Locati M, Mantovani A. Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression. Journal of immunology 177, 7303-7311 (2006).

42. Mao, Q., Wang, L., Tsang, I. Sun, Y., et al. Principal Graph and Structure Learning Based on Reversed Graph Embedding. IEEE Trans. Pattern Anal. Mach. Intell. 37 2227 - 2241 (2016).

43. X Qiu, et al. Reversed graph embedding resolves complex single-cell trajectories. Nature Methods 14, 979-982 (2017).

44. Wu AR, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods 11, 41-46 (2014).

45. Klein U, et al. Transcriptional analysis of the B cell germinal center reaction. Proceedings of the National Academy of Sciences of the United States of America 100, 2639-2644 (2003).

46. Jiang Y, Li Y, Zhu B. T-cell exhaustion in the tumor microenvironment. Cell death & disease 6, e1792 (2015).

47. Pfoertner S, et al. Signatures of human regulatory T cells: an encounter with old friends and new players. Genome biology 7, R54 (2006).

48. Fuertes Marraco SA, Neubert NJ, Verdeil G, Speiser DE. Inhibitory Receptors Beyond T Cell Exhaustion. Frontiers in immunology 6, 310 (2015).

49. Singh JK, Simoes BM, Howell SJ, Farnie G, Clarke RB. Recent advances reveal IL-8 signaling as a potential key to targeting breast cancer stem cells. Breast cancer research : BCR 15, 210 (2013).

50. Amir AD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol 6, 545-552 (2013).

51. Sofia Edin, Maria L. Wikberg, et al. The Distribution of Macrophages with a M1 or M2 Phenotype in Relation to Prognosis and the Molecular Characteristics of Colorectal Cancer. PLoS One 7, e47045 (2012).

52. Gajewski TF, Schreiber H, Fu YX. Innate and adaptive immune cells in the tumor microenvironment. Nature immunology 14, 1014-1022 (2013).

53. Nguyen LT, Ohashi PS. Clinical blockade of PD1 and LAG3--potential mechanisms of action. Nature reviews Immunology 15, 45-56 (2015).

54. Johnston RJ, et al. The immunoreceptor TIGIT regulates antitumor and antiviral CD8(+) T cell effector function. Cancer cell 26, 923-937 (2014).

55. Michael J, et al. Macrophage M1/M2 Polarization Dynamically Adapts to Changes in Cytokine Microenvironments in Cryptococcus neoformans Infection. mBio 4, e00264-13 (2013).

56. Takahisa Y, et al. Tumor-associated macrophages of the M2 phenotype contribute to progression in gastric cancer with peritoneal dissemination. Gastric Cancer 19, 1052–1065 (2016).

57. Lee JH, et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nat Protoc 10, 442-458 (2015).

국문 초록

서론: 암 미세 환경은 종양 세포, 기질 세포 및 면역 세포의 세포 혼합을 비롯한 다양한 구성 요소에 의해 형성된다. 종양 샘플의 단일 세포 전사체 프로파일링은 이질성의 종양 세포 및 주변 기질 세포 혹은 면역 세포의 구분을 가능하게 한다. 침윤성 면역 세포의 정확한 특성 규명은 면역 요법을 위한 새로운 전략에 대한 단서를 제공 할 수 있다.

방법: 총 11 명의 유방암 환자로부터 얻은 515 개의 세포와 4 명의 진행성 위암 (AGC) 환자로부터 얻은 162 개의 세포가 단일 세포 RNA 시퀀싱 (RNA-seq) 을 통해 분석되었다. M1 형 또는 M2 형 대식세포에 대한 단일 세포 전사체는 정상 혈액 유래 단핵 세포로부터 실험적인 분화를 거쳐 생성되었다.

결과: 단일 세포 RNA-seq 데이터에서 유추된 유전자 복제수 변이 (CNA) 패턴은 종양 세포와 비 종양 세포를 구분하였다. 비 종양 세포의 대부분은 면역 세포의 특징을 보였다. 유방암에서는 T 림프구, B 림프구 및 대식세포의 세 가지 면역 세포가 확인되었다. T 림프구는 조절 T 세포나 exhausted T 세포의 표현형을 보이며 면역 억제성 특징을 나타냈다. B 림프구는 두 개의 하위 그룹으로 나뉘어졌는데,

하나는 세포 사멸을 억제하는 미접촉 T 세포 혹은 기억 T 세포 그룹이고, 다른 하나는 활발히 증식하는 B 세포 그룹이다.

AGC 에서 모든 면역 세포는 종양 관련 대식 세포 (TAM)였다. M1 또는 M2 대식세포의 기준 전사체와 비교했을 때, AGC 의 종양 관련 대식 세포에서 M2 편향된 성향이 이질적인 극화 수준으로 관찰되었다. 이와 대조적으로, 유방암 또는 대장암에서 비롯된 종양 관련 대식 세포는 M1 또는 M2 으로 편향된 세포를 모두 나타내었다.

결론: 종합적으로, 단일 세포 전사체 분석은 암 미세 환경에 있는 면역 세포를 세포 유형과 경로 활성화의 이질적인 수준에 따라 세분화하고, 이는 소진된 종양 침윤성 림프구를 표적으로 하거나 항염증성 M2 형 TAM 을 염증성 M1 형으로 전환시키는 방법에 새로운 관점을 제공 할 수 있다. 이 연구는 면역 요법 전략을 개발하기 위해 종양에 침투하는 면역 세포의 특성을 규명하는 데에 있어 단일 세포 RNA 시퀀싱의 능력을 입증한다.

주요어 : 단일 세포, 유방암, 위암, 세포 이질성, 암 미세 환경, 종양 관련 대식 세포

학 번 : 2012-21802