

SUPPLEMENTARY MATERIALS

1. DATA SUMMARY

Breast cancer Dataset a: The fresh frozen breast tissue from Invasive Ductal Carcinoma acquired by 10x Genomics from BioIVT Asterand is used in the spatial gene expression dataset. The tissue was embedded, cryosectioned into 10 μm sections, and mounted onto Visium Gene Expression Slides using Visium Spatial Protocols (CG000240). H&E photos were captured using a Nikon Ti2-E microscope equipped with a color camera, 20X objective (Numerical Aperture: 0.75), 10 ms exposure, and 4.5X gain. The described tissue was annotated for ductal carcinoma *in situ*, lobular carcinoma *in situ*, and Invasive Carcinoma. It also showed AJCC/UICC Stage Group IIA, ER positive, PR negative, and Her2 positive indicators. Following the Visium Spatial Gene Expression Reagent Kits User Guide (CG000239), the library was prepared in the following steps. 3987 Identified spots were found using sequencing on an Illumina NovaSeq 6000 with a sequencing depth of 138226 reads per spot and a 28×120 bp arrangement. The spots showed a median gene count of 5580 per spot and a median UMI count of 18828. The dataset, which was sequenced on Slide V19L29-098 in Area B1 using the T1T2-H10 Dual-Index set, offers valuable molecular information across multiple spatial locations for a thorough analysis and understanding of the disease. It offers comprehensive insights into the spatially resolved gene expression patterns within the breast tissue of patients with invasive ductal carcinoma.

Breast cancer Dataset b: The fresh frozen breast tissue from Invasive Ductal Carcinoma obtained by 10x Genomics from BioIVT Asterand is the subject of this spatial gene expression dataset. Via cryosectioning and embedding, the tissue was divided into 10 μm thick sections that were placed on Visium Gene Expression Slides using the Visium Spatial Protocols (CG000240). Acquiring H&E photos was made easier by imaging with a Nikon Ti2-E microscope that had a color camera, a 20X objective (Numerical Aperture: 0.75), a 10 ms exposure, and 4.5X gain. The analyzed tissue had annotations for invasive carcinoma, lobular carcinoma *in situ*, and DUCC carcinoma *in situ*. It also showed signs of ER positive, PR negative, and Her2 positive cancer. It was also in AJCC/UICC Stage Group IIA. As per the instructions provided in the Visium Spatial Gene Expression Reagent Kits User Guide (CG000239), the Visium Spatial Gene Expression library was assembled. A dataset with a sequencing depth of 149800 reads per spot using a 28×120 bp arrangement was obtained by using Illumina NovaSeq 6000 sequencing after later. A median gene count of 5394 per spot and a median UMI count of 17531 were found in 3813 identified spots, according to the analysis. This dataset offers significant molecular information across multiple spatial locations for in-depth investigation and comprehension of the disease characteristics. It was sequenced using the T1T2-F10 Dual-Index set on Slide V19L29-097 in Area B1. This dataset offers detailed insights into spatially resolved gene expression within the breast tissue of patients with invasive ductal carcinoma.

Breast cancer Dataset c: Fresh frozen Invasive Ductal Carcinoma breast tissue obtained from BioIVT Asterand by 10x Genomics is analyzed for the spatial gene expression dataset. After tissue was prepared using the Visium Spatial Protocols, it was embedded and cryosectioned into 10 μm sections that were then deposited onto Visium Gene Expression Slides. imaging with a Nikon Ti2-E microscope that produced H&E images with a 4.5X gain, a 20X objective, and a numerical aperture of 0.75 over a 10-millisecond exposure. The tissue under investigation was classified as invasive carcinoma, lobular carcinoma *in situ*, and ductal carcinoma *in situ*. It was identified as belonging to AJCC/UICC Stage Group IIA and had ER

positive, PR negative, and Her2 positive indicators. The user guide for the Visium Spatial Gene Expression Reagent Kits (CG000239) was followed during library preparation. A median of 20762 UMIs and 6026 genes were found in 3798 identified spots using an Illumina NovaSeq 6000 sequencing configuration at a depth of 150392 reads per spot. These insightful results were obtained by a 28×120 bp arrangement. The dataset offers a thorough overview of the spatial gene expression profiles inside the breast tissue impacted by invasive ductal carcinoma. It was processed using a T1T2-F10 Dual-Index set on Slide V19L29-097 in Area B1. This set of data contains pertinent molecular information across several spatial regions.

Breast cancer Dataset d: Fresh frozen breast tissue from Invasive Ductal Carcinoma, obtained from BioIVT Asterand by 10x Genomics, is used in the spatial gene expression dataset. The tissue is processed using the Visium Spatial Protocols (CG000240), which include tissue embedding and cryosectioning into 10 μ m sections that are then placed onto Visium Gene Expression Slides. H&E photos were obtained with a Nikon Ti2-E microscope that was outfitted with a color camera, 20X objective (Numerical Aperture: 0.75), 10 ms exposure, and 4.5X gain. The identified tissue had ER positive, PR negative, and Her2 positive markers, and it was annotated for invasive carcinoma, lobular carcinoma *in situ*, and ductal carcinoma *in situ*. It also showed AJCC/UICC Stage Group IIA. The Visium Spatial Gene Expression Reagent Kits User Guide (CG000239) was followed in the preparation of the Visium Spatial Gene Expression library. Using a 28×120 bp arrangement, sequencing on the Illumina NovaSeq 6000 produced a dataset with a sequencing depth of 137262 reads per spot. 4015 detected spots with a median gene count of 5100 and a median UMI count of 16502 were found by the study. This dataset provides important insights into the spatially resolved gene expression profiles within the breast tissue of invasive ductal carcinoma (IDC), having been sequenced using the T1T2-H10 Dual-Index set on Slide V19L29-098 in Area B1. This information is crucial for a thorough investigation of the disease's features as it spans multiple spatial locations.

Colorectal cancer Dataset α : The FFPE human colorectal cancer tissue used in the spatial gene expression dataset was acquired from Discovery Life Sciences and processed by 10x Genomics utilizing Visium CytAssist Spatial Gene Expression techniques. The tissue was two years old in the FFPE block when it was identified as colorectal cancer, adenocarcinoma, with Stage II-A in the AJCC system and T3 in the TNM system. Following the Deparaffinization, H&E Staining, Imaging & Decrosslinking Demonstrated Protocol (CG000520), a 5 μ m tissue section was sectioned in accordance with the Visium CytAssist Spatial Gene Expression for FFPE – Tissue Preparation Guide (CG000518). Using an Olympus VS200 Slide Scanning Microscope, the colorectal cancer was observed, exhibiting a desmoplastic response and invading tumor areas with a high tumor stroma content. After processing the tissue in accordance with the Visium CytAssist Spatial Gene Expression Reagent Kits User Guide (CG000495), 11 mm capture regions were produced using the Visium CytAssist Spatial Gene Expression Slide v2. According to the Visium CytAssist Spatial Gene Expression for FFPE User Guide (CG000495), library preparation was done. A high-depth dataset with 911044482 reads and 46.2 % saturation was obtained by sequencing on an Illumina NovaSeq 6000 using flow cells HJWFGDSX5 and HJW2WDSX5 (lane 1–4). After analyzing Slide V52Y10-310 in Area A1, extensive insights into the spatially resolved gene expression in FFPE human colorectal cancer tissue were obtained. These insights are essential for comprehending the molecular landscape and characteristics of the tumor microenvironment. The analysis revealed 9,080 spots with a median UMI count of 49282 and a median gene count of 9560 per spot.

Colorectal cancer Dataset b: Large intestinal colorectal cancer, represented by a 5 μ m segment from Block 1281585B, is the subject of the spatial gene expression dataset, which is based on FFPE human intestinal cancer tissue that was received from BioIVT Asterand human tissue specimens. With a Brightfield camera set to 20x magnification, the tissue section was photographed. Using the Visium Spatial Gene Expression Reagent Kits for FFPE User Guide (CG000407 Rev A), libraries were created. Using flow cell HCCVHDSX2

(lanes 1–2) and a sequencing depth of 77288 reads per spot, the Illumina NovaSeq 6000 was used for the subsequent sequencing. The configuration included 28 bp read 1 (16 bp Visium spatial barcode, 12 bp UMI), 50 bp read 2 (transcript), and 10 bp i7 and i5 sample barcodes in the Dual-Index set SI-TS-B1. 2660 spots with a median UMI count of 25312 and a median gene count of 7438 were found on Slide V10L13-021 in Area B1. This dataset enables a thorough investigation of the molecular markers and tumor microenvironment features connected to FFPE Human Intestinal Cancer. It also sheds light on the spatially resolved gene expression patterns within this kind of cancer.

Colorectal cancer Dataset c: Fresh frozen invasive adenocarcinoma of the large intestine tissue obtained from BioIVT Asterand and processed by 10X Genomics utilizing Visium Spatial Protocols and Methanol Fixation, H&E Staining & Imaging (CG000160) constitute the subject matter of the spatial gene expression dataset. A Nikon Eclipse Ti2-E microscope with a color camera, 10X objective (Numerical Aperture: 0.45), and a 20 ms exposure time was used to image the tissue, which was identified as AJCC/UICC T4aN0M0, Stage Group IIB. In accordance with the Visium Spatial Reagent Kits User Guide (CG000239 Rev D), the Visium Gene Expression library (T1T2-E8) was created. The Human Gene Signature Panel (PN-2000285) was specially used to build targeted libraries in accordance with the Targeted Gene Expression - Spatial User Guide (CG000377). Using a paired-end (28 X 90) arrangement, sequencing on the Illumina NovaSeq 6000 with flow cell HHW57DSXY (lane 4) produced a mean of 16795 reads per cell, allowing 88.2 % of reads to accurately map to the targeted transcriptome. 3138 spots were found on Slide V10A13-206 in Area C1, indicating a total of 1134 targeted genes with a median of 351 targeted genes and 635 targeted UMIs per spot. This dataset provides comprehensive information about the expression of targeted genes and aids in understanding the molecular subtleties of invasive adenocarcinoma of the large intestine by providing detailed spatial insights into the expression of a particular gene signature panel within the cancer type.

Lung cancer Dataset a: The dataset consists of 5' gene expression data obtained using Cell Ranger 2.2.0 for Single Cell Immune Profiling from a recent surgical resection of a squamous non-small cell lung cancer (NSCLC) tumor. Using the V(D)J Reagent Kits User Guide (CG000086 Rev C), the cDNA from this tumor was used to create particular libraries for T cell receptor (TCR) and B cell immunoglobulin (Ig) profiling. After loading about 15000 cells, it was predicted that 9000 viable cells would be recovered based on cell viability and capture efficiency, leading to the final detection of 7802 cells. With a median of 1442 genes and 4197 UMI counts, each cell provided an extensive picture of gene expression. About 48000 read pairs per cell were produced during sequencing on an Illumina NovaSeq 6000 using NovaSeq software v1.2. The read configuration was 150 bp read1 (with a 16 bp Chromium barcode and 10 bp UMI), 150 bp read2 (transcript), and 8 bp i7 sample barcode. After the analysis was carried out with `--force-cells = 7802`, 7,802 cells were definitively identified. This allowed for a thorough investigation of the single-cell transcriptome landscape in this NSCLC tumor in order to fully profile the immune system and comprehend tumor heterogeneity at the molecular level.

Lung cancer Dataset b: The Single Cell Immune Profiling Dataset using Cell Ranger 5.0.0 focuses on a fresh surgical resection of a squamous non-small cell lung carcinoma (NSCLC) tumor obtained by 10x Genomics. About fifteen thousand of the processed cells were expected to be recovered after live cell sorting, resulting in about nine thousand viable cells. The construction of gene expression, T cell receptor (TCR) amplified, and B cell receptor (BCR) amplified libraries was done in accordance with the Single Cell V(D)J Reagent Kits User Guide (CG000086 Rev C). Illumina NovaSeq 6000 was used for sequencing. Paired-end sequencing (150 × 150) was performed on the libraries using single-indexed reads for the Gene Expression, TCR Amplified, and BCR Amplified datasets. Each read set consisted of two parts: Read 1 (150 cycles) which recorded the transcript and i7 index (8 cycles), and Read 2 (150 cycles) which recorded additional transcript data. To ensure consistent cell identification across different library types within the same sample, the Gene

Expression, Cell Surface Protein/Immune Receptor Mapping, TCR Amplified, and BCR Amplified libraries were combined using the Cell Ranger multi process. The thorough investigation offers an in-depth perspective of the tumor microenvironment, making it easier to characterize immune cells and the gene expression patterns inside the NSCLC tumor.

Lung cancer Dataset c: This dataset contains spatial gene expression data obtained from a human lung cancer tissue sample, specifically identified as neuroendocrine carcinoma, at stage IB according to the AJCC staging system. The tissue was acquired from Avaden Biosciences and processed using the Visium technology by 10x Genomics. A 5 μ m thick piece of the tissue was cut and put on a regular glass slide in order to prepare it for analysis. The tissue was then stained in accordance with the guidelines provided for the production of FFPE tissue. After that, the stained tissue segment was run through the Visium CytAssist apparatus in order to transfer analytes onto the 11 mm-capacitor Visium CytAssist Spatial Gene Expression Slide v2. The diagnosis, staging information (TNM system), age of the FFPE block (less than 1 year), and DV200 value, which indicates RNA quality, are important characteristics regarding the tissue. Notably, features including vascular invasion, squeezed bleeding inside the tumor, and a visible stromal boundary encircling the solid tumor region are revealed by the H&E imaging of the tissue. The Visium CytAssist Spatial Gene Expression Reagent Kits for FFPE User Guide was followed in the preparation of libraries for sequencing. Employing specialized flow cells (HJWFGDSX5 and HJW2WDSX5) spanning lanes 1–4, an Illumina NovaSeq 6000 instrument was used to sequence the spatially resolved transcriptome. A 45.9 % saturation rate was attained with 844866053 reads in the sequencing depth A 28 bp read 1 with a 16 bp Visium spatial barcode and a 12 bp UMI is included in the sequencing configuration specifications. A 90 bp read 2 with transcript data follows. Along with the dual-index set SI-TS-B3, 10bp i7 and i5 sample barcodes were also employed. Key cell metrics are supplied, such as the number of spots found under tissue (6195), the median UMI counts per spot (66268), and the median genes per spot (10087). The dataset is linked to a particular slide (V52Y10-286) and area (B1). These metrics provide information about the lung cancer tissue sample's spatial gene expression profile.

Ovarian cancer Dataset a: The Human Ovarian Cancer spatial gene expression dataset obtained using Space Ranger 2.0.0 from 10x Genomics features FFPE tissue blocks sourced from Avaden Biosciences. After sectioning the tissue using Visium CytAssist Spatial Gene Expression for FFPE procedures (CG000518), the tissue—which was identified as Ovarian Serous Carcinoma, High Grade—was placed onto a regular glass slide. Eleven months of room temperature storage elapsed after the initial H&E staining. Following storage, sections underwent protocol-compliant reimage, deparaffinization, dehydration, and decrosslinking (CG000520). The tissue sections were prepared and analytes were transferred to a Visium CytAssist Spatial Gene Expression slide with an 11 mm Capture Area utilizing the Visium CytAssist equipment. The imaging process used an Olympus slide scanning microscope with a 20X UPLXAPO objective that was adjusted to a certain degree. The Visium CytAssist Spatial Gene Expression Reagent Kits for FFPE User Guide (CG000495) was followed in the preparation of the libraries. Key metrics were obtained via sequencing on an Illumina NovaSeq with a depth of 46109 reads per spot: 4674 spots were found under tissue, with a median of 8792 genes and 32048 UMI counts per spot. This dataset allows for a thorough investigation of the molecular landscape within the context of FFPE tissue by providing precise insights into the spatial gene expression profiles within Ovarian Serous Carcinoma, High Grade.

Ovarian cancer Dataset b: The spatial gene expression dataset for Human Ovarian Cancer, captured using Space Ranger 1.2.0, involves fresh frozen Endometrial Adenocarcinoma of the ovary tissue from BioIVT Asterand. Following the Visium Spatial Protocols, the tissue was cryosectioned, and the resulting 10 μ m slices were mounted on Visium Gene Expression slides. In accordance with Methanol Fixation, Immunofluorescence Staining & Imaging protocols for Visium Spatial analysis, antibody staining with Alexa Fluor 488 anti-Cytokeratin (pan reactive) and Alexa Fluor 647 anti-human CD45, together with

DAPI for nuclei, was carried out. The tissue was classified as Stage Group I, AJCC/UICC T1N0M0. Four channels were imaged using a Nikon Eclipse Ti2E microscope operating at 20X: DAPI, Cytokeratin, Fiducial Frame, and CD45. Following standard protocols (CG000239 Rev D) for the production of the Visium Gene Expression library (T1T2-H6), sequencing on an Illumina NovaSeq 6000 (flow cell HHYWHDSXY) produced paired-end reads (28 X 90) with an average of 80586 reads per cell. The dataset showed 3493 spots that were discovered, with a median of 3464 genes and 8095 UMI counts per spot. This dataset facilitates detailed investigation of spatially resolved gene expression in ovarian cancer tissue, allowing for a thorough examination of gene expression patterns in connection to immune markers such as CD45 and cellular localization.

Ovarian cancer Dataset c: The spatial gene expression dataset from Human Ovarian Cancer (FFPE) using Space Ranger 1.3.0 features a 5 μ m section from serous papillary carcinoma of the human ovary obtained from iSpecimen. Using a Brightfield camera set to 20x magnification, the tissue block—which was revealed to be serous papillary carcinoma, an ovarian cancer—was photographed. Flow cell HCCVHDSX2 was used to perform sequencing over lanes 1–4 on an Illumina NovaSeq 6000 platform. Sequencing depth of 63836 reads/spot was attained utilizing a set-up of 28 bp for read 1 (which consisted of a 16 bp Visium spatial barcode and a 12 bp UMI) and 75 bp for read 2 (transcript), in addition to 10 bp indices for the i7 and i5 sample barcodes using Dual-Index set SI-TS-H1. V10L13-020 in Area D1 was the precise slide that was used. Important metrics that were found include 3455 spots that were found under tissue, with a median of 7621 genes per spot and a median UMI count of 28927 per spot. This dataset provides critical insights into the molecular landscape of serous papillary carcinoma in the setting of FFPE tissue by enabling a thorough investigation of the spatially resolved gene expression patterns inside the cancer.

2. QUALITY CONTROL AND FILTERING

In order to determine the datasets overall quality, we carefully analyzed them throughout the quality control phase. We used a variety of methods to assess the cells inside the Seurat Objects. Initially, we thoroughly examined the metadata linked to each dataset, comprising four columns: cell names, dataset identity, RNA molecule count, and gene count per cell. This ensured data consistency and reliability. To assess mitochondrial percentages, we identified mitochondrial genes via the ‘MT-’ pattern and computed their proportion in each cell. This added a column to the metadata displaying mitochondrial gene percentages. Elevated mitochondrial gene expression may indicate low-quality cells or technical issues. By calculating mitochondrial percentages, we excluded low-quality cells, focusing solely on high-quality ones. Violin plots (Fig. S2, Fig. S3) and scatter plots (Fig. S4, Fig. S5) were utilized for comprehensive evaluation of gene expression distribution and dataset quality [1].

A high-quality dataset should exhibit consistent and uniform gene expression patterns across cells, particularly for highly expressed genes. Tight, well-defined violin plots with distinct peaks (Fig. S2, Fig. S3) evidence this indicating minimal technical noise. Any significant fluctuations or expansions in expression patterns may suggest system glitches or batch effects. Additionally, a narrow violin plot with low density for mitochondrial genes signifies acceptable data quality and healthy cells with low mitochondrial gene expression. Outliers or a subset of cells with elevated mitochondrial gene expression levels may indicate potential issues such as cell stress, low viability, or contamination during sample handling. Increased percentages of mitochondrial gene expression could further suggest compromised cell health. In single-cell RNA sequencing (scRNA-seq) data, there is an expected correlation between RNA molecule counts (indicating sequencing depth or cell size) and expressed genes (reflecting cellular complexity). Deviations from this correlation, detectable via scatter plots (Fig. S4, Fig. S5), may indicate doubts, technical errors, or low-quality cells. Insights from scatter plots inform criteria for excluding low-quality cells and adjusting data pre-processing steps. This rigorous quality control process ensures high-quality data for reliable research outcomes.

3. DATA NORMALIZATION

We employed normalization techniques to enhance distribution symmetry and mitigate the impact of exceptionally high expression levels in gene expression data. Specifically, log-normalization, as outlined by [2], was applied to the dataset to ensure comparability and stabilize variance. This transformation condensed values, narrowing the dynamic expression range, facilitating inter-cell expression comparisons. Log normalization effectively addressed expression level discrepancies, ensuring proper data scaling for subsequent analysis. The log-normalization formula can be expressed mathematically as:

$$\text{log-normalized expression} = \log \left(\frac{\text{row count of gene } i \text{ in cell } j}{\text{sum of counts in cell } j} \times \text{scale factor} + 1 \right)$$

where the *row count of gene i in cell j* is the original count of gene i in a particular cell j ; the *sum of counts in cell j* is the total count of all the genes in cell j , *scale factor* was employed for extra normalization to account for variations in library size or sequencing depth between cells. We fixed the *scale factor* as 10000. The log function is the natural logarithm.

4. HIGH VARIABLE FEATURES EXTRACTION

Following dataset normalization, we identified the top 2000 highly variable features (Fig. S6, Fig. S7). This process is akin to highlighting the most impactful traits within a vast dataset, streamlining analysis by focusing on genes exhibiting significant cell-to-cell differences. These highly variable genes are pivotal for discerning diverse cell types and facilitate efficient downstream analyses like PCA, clustering, and UMAP visualization. Subsequently, we applied variance stabilizing transformation (VST) method to pinpoint genes exhibiting notable expression level variations across cells post log-normalization of scRNA-seq data [3]. Following are some of the basic mathematical criteria that we used for feature extraction:

Mean Normalized Expression (μ_i): The average log-normalized expression value for each gene i over all cells has been calculated to determine the mean expression of that gene.

$$\mu_i = \frac{1}{N} \sum_{j=1}^N \text{log-normalized expression}_{ij},$$

where N is the total number of cells; μ_i is the mean normalized expression in gene i .

Population Variance (σ^2): The variation of each gene's expression levels across cells is computed using the variance equation utilized in the context of VST. The following is the general formula for calculating variance:

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^N \left(\text{log-normalized expression}_{ij} - \mu_i \right)^2$$

where σ^2 represents the population variance of gene i .

5. DATA SCALING

In order to analyse expression complexity and stabilize variance, we initially employed log-normalization. Next, to mitigate technical variances, we uniformly scaled for expression normalization within cells post log-normalization [2]. This ensured consistent gene expression across cells, vital for accurate clustering analysis. Scaling improved gene contribution consistency, unveiling biologically significant patterns. Our aim was to minimize technical noise and accentuate true biological expression variations. The combined use of scaling and log-normalization enhanced data clarity, facilitating precise analysis and deeper insight into gene-cell relationships. Thus, our systematic approach ensured optimal data quality for robust biological inferences. We employed Seurat's 'ScaleData' function for scaling scRNA-seq data

within cells, adhering to Stuart et al.’s methodology [2]. This function standardizes gene expression levels within cells by mean-centering and scaling by standard deviation, reducing technical variability and enhancing downstream analysis precision.

Compute the Mean and Standard Deviation. Calculate the mean (μ) and standard deviation (σ) of gene expression values:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i, \quad \sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2},$$

where n denotes the total number of cells, x_i denotes the expression level of the gene in cell i .

Scaling the Data. For a gene expression value x_i in a specific cell, the scaled value (x_{scaled}) is calculated as follows:

$$x_{\text{scaled}} = \frac{x_i - \mu}{\sigma}.$$

We standardized the data by setting its standard deviation to one and centering it at zero. To maintain consistency and control without affecting integrity, we capped the scaled value at 10. For computational efficiency, we implemented scaling in blocks, with a block size of 1000 features. Blocks smaller than 3000 cells were exempt from scaling, optimizing the process for larger, more significant data segments.

6. LINEAR DIMENSIONALITY REDUCTION

In our analysis, we identified the top 2000 variable features or genes from our single-cell RNA sequencing data using Seurat. These genes demonstrate the highest variability between cells, serving as crucial markers for diverse biological properties [4]. Subsequently, we employed Principal Component Analysis (PCA) to reduce the dataset’s dimensionality while preserving its essential biological information. PCA generated uncorrelated variables, capturing the primary sources of variance in the dataset. This reduction facilitated efficient computations, enhanced data visualization, and enabled the identification of unique cell subgroups based on gene expression patterns.

The optimal number of principal components to retain for subsequent analyses, such as clustering, visualization, or additional dimensionality reduction methods like UMAP, was determined using an elbow plot (Fig. S8, Fig. S9). This approach balances reducing data dimensionality for effective analysis and visualization while capturing the most significant sources of variation in the dataset. Employing the elbow plot was crucial in our PCA analysis to determine the appropriate number of principal components to retain [5], enabling us to strike a balance between dimensionality reductions and preserving important variance. The elbow plot visually depicted the diminishing standard deviation with each additional principal component (Fig. S8, Fig. S9). Our focus was on identifying the “elbow” of this curve, representing the point where the standard deviation decreased at a significantly slower rate.

7. CLUSTERING

Following the application of principal component analysis (PCA) to condense our single-cell RNA sequencing data and identifying the most informative principal components, it was crucial to categorize cells based on similar expression profiles. Utilizing clustering, cells were grouped into distinct clusters according to gene expression patterns, aiming to identify inherent cellular subgroups or clusters within the dataset [6]. This process, depicted in Fig. S10 and Fig. S11, involved mathematical computations involving following technical steps.

Search for K-Nearest Neighbors (KNN). We used the KNN algorithm to find the 20 closest neighbors for each cell in the smaller space. This method allows the accurate identification of neighborhood structures and cell-cell interactions in the single-cell RNA sequencing dataset by using distance metrics.

Graph Construction. In the graph, a node represents every cell in the dataset. Nodes form connections (edges) with each other according to their neighbor relationships or closeness. To create a graph structure where nodes represent cells and edges denote KNN interactions, we can connect edges between a cell (node) and its K -nearest neighbors. This graph representation forms the basis for analyzing cell relationships, clustering, and visualizing the structure of the single-cell RNA sequencing data.

Compute Shared Nearest Neighbor (SNN). Using the graph G , we can find the neighbors of cells i and j in each neighborhood and ascertain how many neighbor cells i and j have in common within each of their respective neighborhoods. The similarity or proximity between cells i and j is determined by counting the number of common neighbors between them. Stronger or more important interactions between cells i and j are indicated by higher shared neighbor counts. We stipulated that pairs of cells are considered to have strong neighborhood interactions if their common neighbor count is greater than or equal to 0.0667. Stronger relationships between cells within the SNN graph can be highlighted by removing any shared neighbor count below this cutoff point from additional analysis or classifying it as less significant.

Approximate Nearest Neighbor (ANN) Algorithm. In the realm of high-dimensional datasets like those in single-cell RNA sequencing (scRNA-seq) analysis, the approximate nearest neighbor (ANN) method serves to expedite the search for closest neighbors while mitigating computational complexity. This approach involves generating multiple trees from the dataset through the ANN algorithm, such as Annoy, to produce an approximate solution instead of exact computations. In our case, 50 trees were constructed to organize the data for swifter searches. Utilizing Euclidean distance metrics, the algorithm efficiently traverses these tree structures without exhaustively assessing every point in the high-dimensional space. It constructs a graph-based representation of scRNA-seq data, enabling the identification of cellular interactions and facilitating subsequent analyses. Following the creation of the neighborhood graph, cells were categorized into distinct clusters based on their proximity in the reduced-dimensional space. We employed the Louvain algorithm as a clustering technique. The Louvain algorithm finds cohesive communities or cell clusters by optimizing the network structure. The goal of clustering algorithms is to maximize modularity, which is a metric for the caliber of cluster assignments made inside the network. The amount of intra-cluster connections in relation to inter-cluster connections is measured by modularity. Following mathematical equation allows calculation of modularity score,

$$Q = \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j),$$

where Q is the modularity score that expresses how well the clustering is done; A_{ij} is an adjacency matrix entry showing the relationship between nodes i and j ; k_i , k_j are the node degrees representing the total of the edge weights attached to nodes i and j ; m is the total of the network's edge weights; and $\delta(c_i, c_j)$ is the Kronecker delta function, equals 1 if nodes i and j are in the same community, otherwise 0. Initially, each node (cell) is assigned to its own community, with nodes shifted across communities iteratively to maximize the modularity score. This involves relocating nodes to the closest community to optimize the gain in modularity. The communities from the first stage form nodes within a higher-level network, which undergo iterative improvement of community structures at various levels. As the algorithm processes cells, cluster assignments are altered to maximize modularity, with cells reassigned between clusters based on modularity scores to enhance the network's overall structure. Cells are then grouped according to optimal clustering outcomes, and each cell is tagged or given a cluster identifier to facilitate later analysis and visualization (Fig. S10, Fig. S11). We employed the resolution argument as a tuning parameter to adjust cluster identification granularity [6]. For each dataset, the resolution parameter is adjusted

individually to produce eight distinct clusters in each set (Table S3). This adjustment aims to achieve a more comprehensive data partitioning for a deeper examination of both datasets.

We determined each data point's silhouette score in order to assess the quality of the clusters we created. The silhouette score measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The values range from -1 to 1 , where a score closer to 1 indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters. Conversely, a score near -1 suggests that the object might be better suited to a neighboring cluster. To get an overall understanding of the clustering quality, we calculated the mean silhouette score for all data points. Following the completion of Uniform Manifold Approximation and Projection (UMAP), we took the subsequent actions to assess the clustering quality using silhouette scores:

UMAP Embeddings. This matrix shows dimensions in the reduced UMAP space as columns and rows, respectively, representing individual data points. Assume that this matrix contains D dimensions and N data points.

Cluster Information. Each data point's allocated cluster labels are contained in this section. Assume that there are K clusters overall.

Silhouette Score Calculation. The pairwise distance matrix is calculated based on the UMAP embeddings. This distance matrix represents the distances between each pair of data points. Let $d(i, j)$ represent the distance between data points i and j . For each data point i in a cluster C_i , calculate the average distance to all other points within the same cluster C_i . The average intra-cluster distance for data point i is denoted by $a(i)$, and it's computed as the average of distances between i and all other points in the same cluster:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j),$$

where $|C_i|$ represents the number of data points in cluster C_i .

For each data point i , then calculated the average distance from that point to all points in the nearest neighboring cluster (i.e., the cluster other than the one to which i belongs) and Find the minimum average distance from data point i to points in other clusters. Let C_k represent the nearest neighboring cluster to C_i . The average nearest-cluster distance for data point i is denoted by $b(i)$, and it's computed as:

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j),$$

here $|C_k|$ represents the number of data points in the nearest neighboring cluster C_k to the cluster C_i that i belongs to. The silhouette score for each data point i is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}.$$

The silhouette score $s(i)$ lies in the range of $[-1, 1]$. A higher silhouette score indicates that the data point is well-clustered (i.e., closer to its own cluster's points and farther from points in neighboring clusters).

Mean Silhouette Score. By averaging each data point's individual silhouette score, the mean silhouette score for all data points is determined:

$$\text{Mean Silhouette Score} = \frac{1}{N} \sum_{i=1}^N s(i)$$

This provides an overall assessment of the clusters' quality and separation. Greater mean silhouette scores signify more precisely defined clusters with suitable spacing between them (Table S4).

8. FIGURES AND TABLES

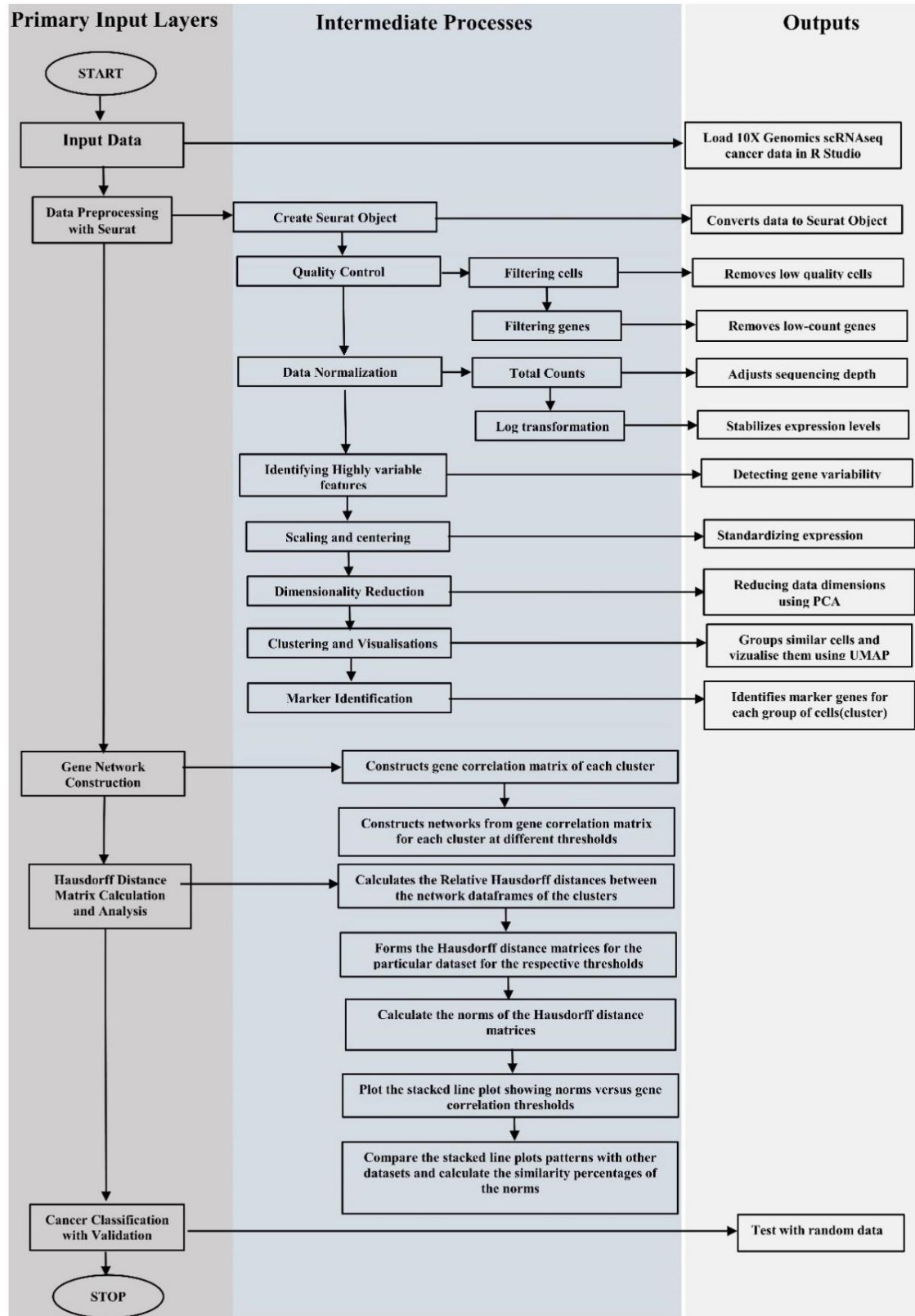


Fig. S1. Cancer classification workflow. It consists of three interconnected layers. The input layer provides the initial data and an overview of the methodology, the intermediate layer details the key methodological steps, and the output layer presents the results generated through the cancer classification process.

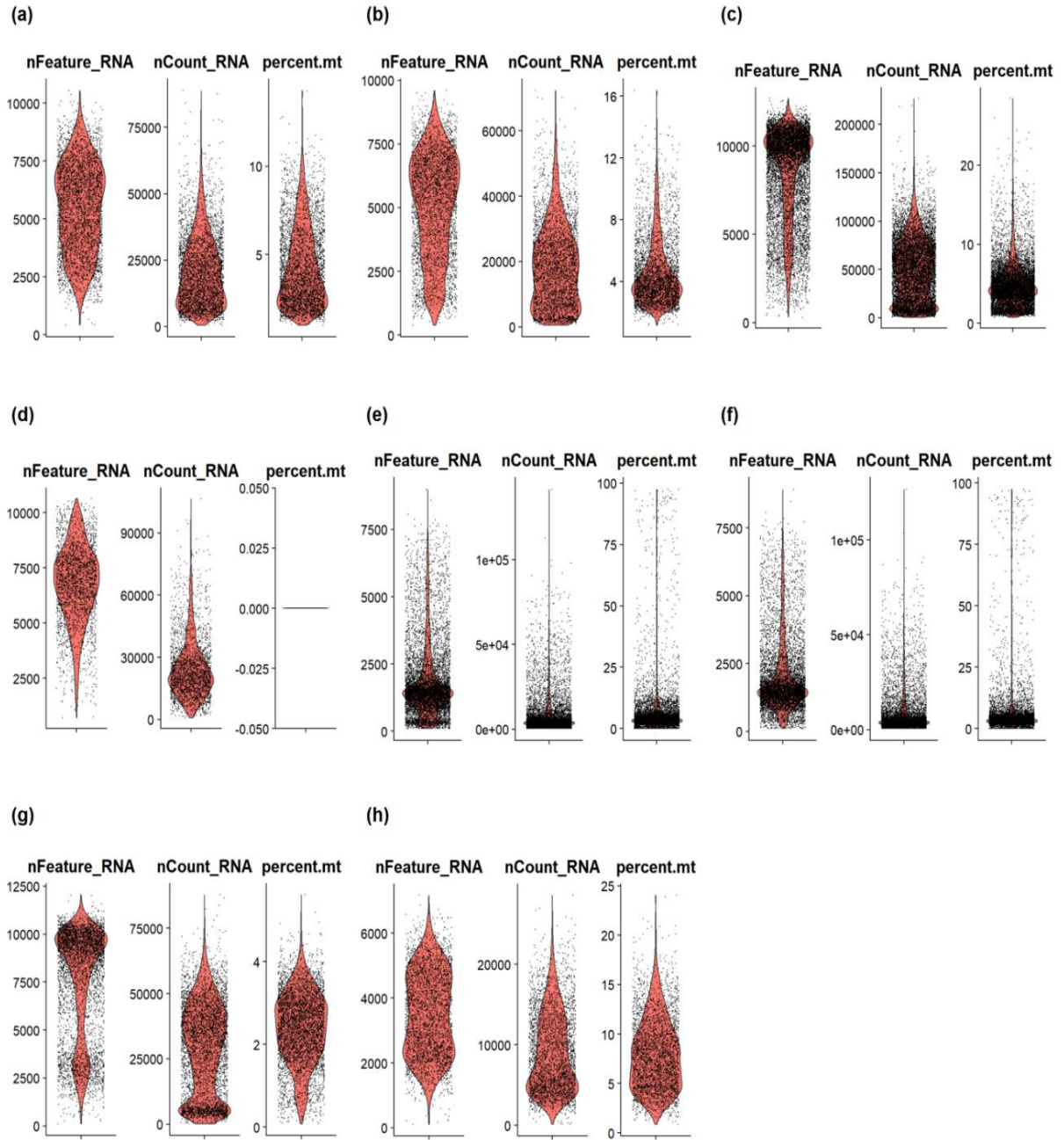


Fig. S2. Violin plots showing the distribution of data described in Table S1. Eight different single-cell RNA sequencing (scRNA-seq) datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression), representing four different types of cancer in human: (a) breast cancer Dataset a, (b) breast cancer Dataset b, (c) colorectal cancer Dataset a, (d) colorectal cancer Dataset b, (e) lung cancer Dataset a, (f) lung cancer Dataset b, (g) ovarian cancer Dataset a, and (h) ovarian cancer Dataset b. These datasets are the training datasets. Tight, well-defined violin plots with distinct peaks indicate minimal technical noise. Any significant fluctuations or expansions in expression patterns may suggest system glitches or batch effects. Moreover, a narrow violin plot with low density for mitochondrial genes signifies acceptable data quality and healthy cells with low mitochondrial gene expression.

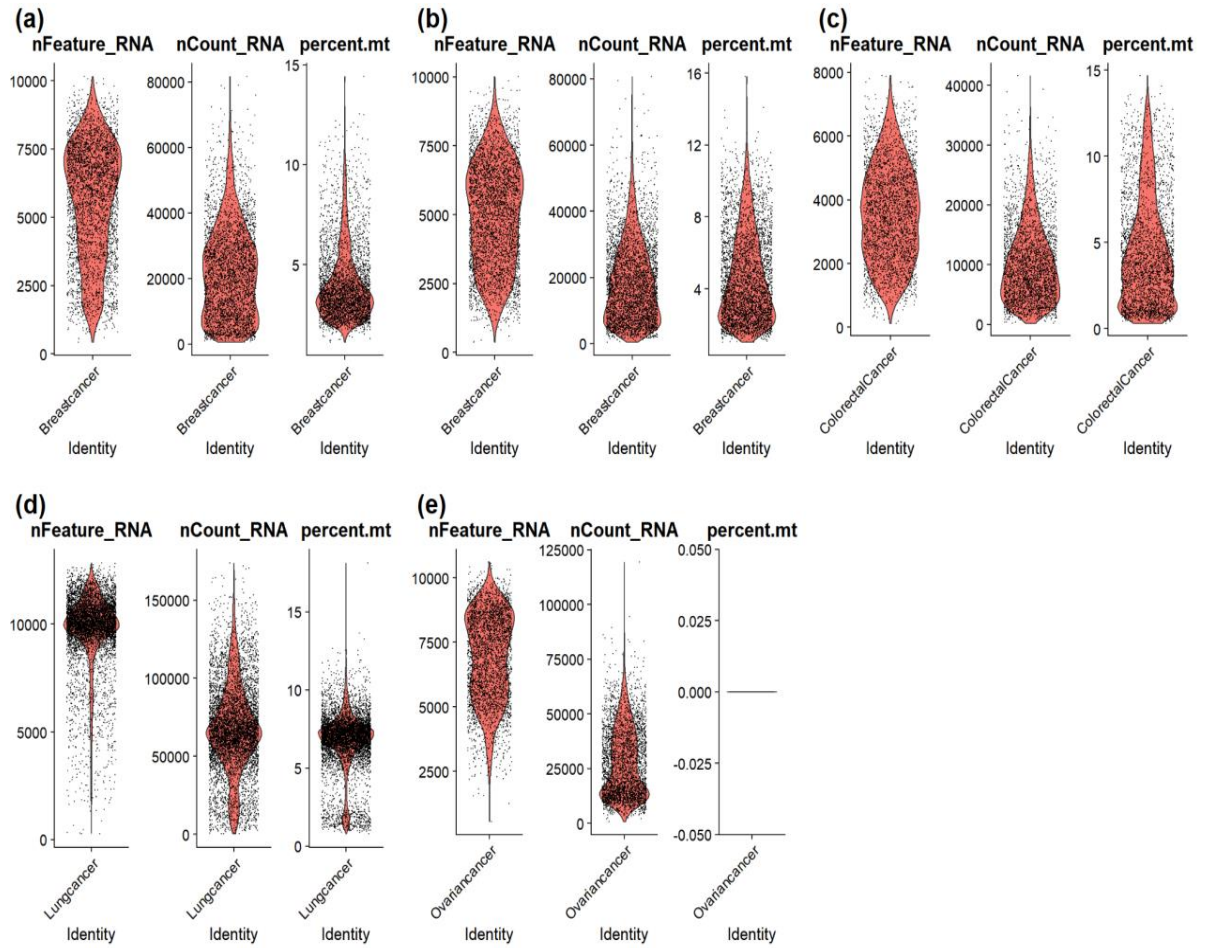


Fig. S3. Violin plots showing the distribution of data described in Table S1. Five different single-cell RNA sequencing (scRNA-seq) filtered cancer datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression): **(a)** breast cancer Dataset c, **(b)** breast cancer Dataset d, **(c)** colorectal cancer Dataset c, **(d)** lung cancer Dataset c, **(e)** ovarian cancer Dataset c. These datasets are the validation datasets. To verify if our methodology correctly classifies them according to their categories, we treated these validation datasets as unknowns even though we knew their cancer types. Tight, well-defined violin plots with distinct peaks are indicating minimal technical noise.

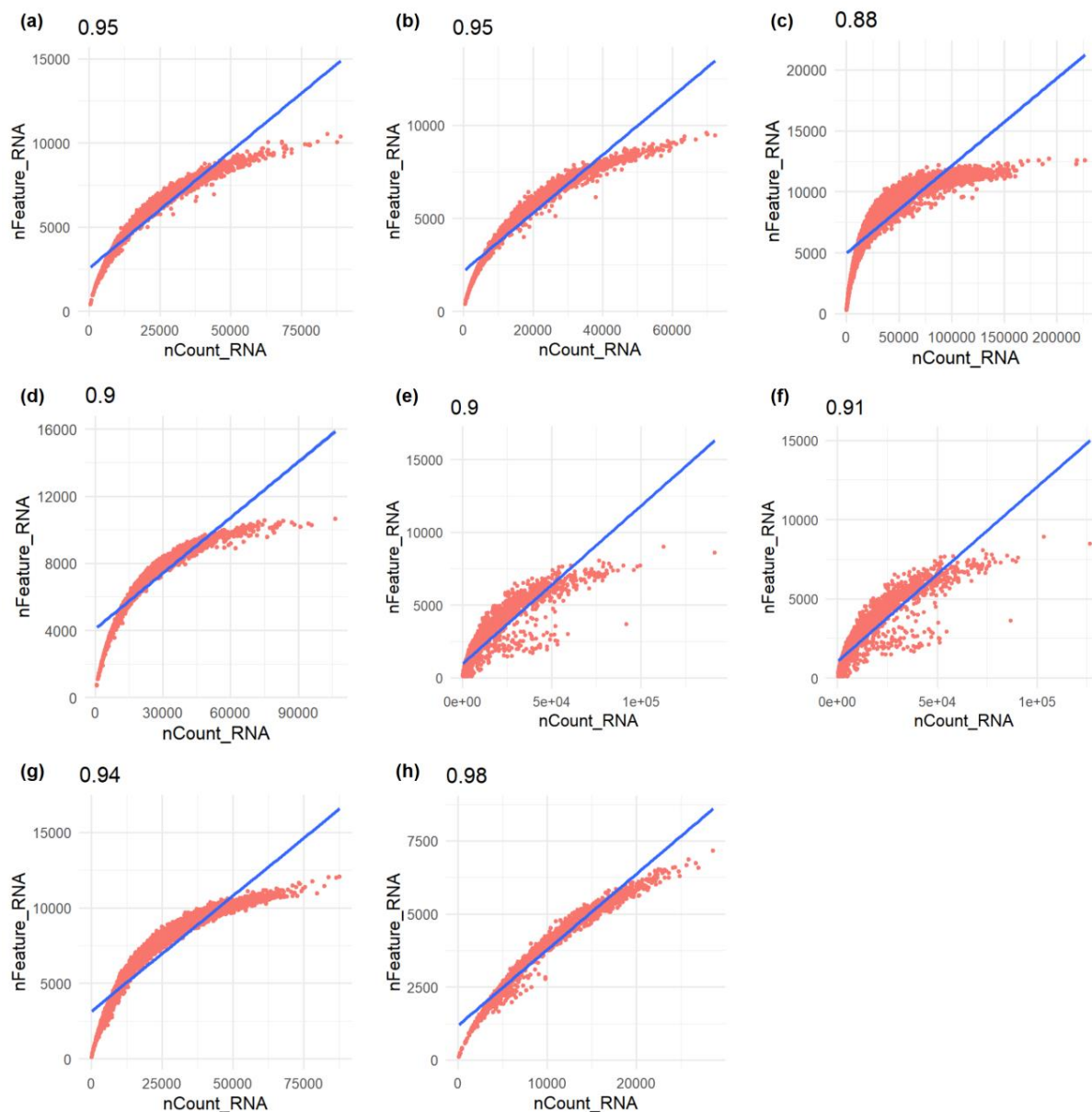


Fig. S4. Scatter plots illustrating the distribution of data described in Table S1. Eight different single-cell RNA sequencing (scRNA-seq) datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression), representing four different types of cancer in Human: (a) breast cancer Dataset a, (b) breast cancer Dataset b, (c) colorectal cancer Dataset a, (d) colorectal cancer Dataset b, (e) lung cancer Dataset a, (f) lung cancer Dataset b, (g) ovarian cancer Dataset a and (h) ovarian cancer Dataset b. These datasets are the training datasets. RNA count typically correlates with expressed genes in scRNA-seq. Deviations from this correlation, detectable via scatter plots may indicate doublets, technical errors, or low-quality cells. Insights from scatter plots inform criteria for excluding low-quality cells and adjusting data preprocessing steps.

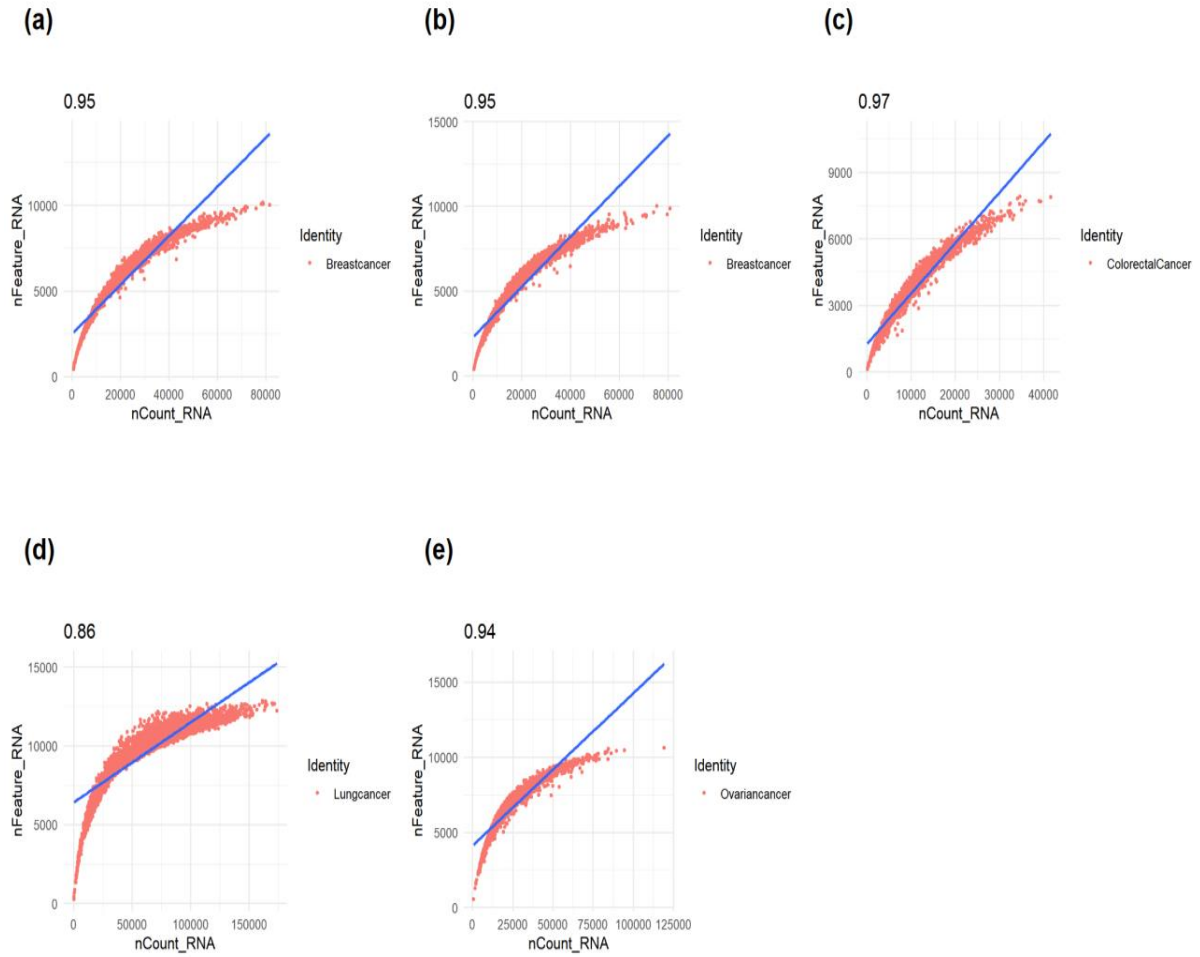


Fig. S5. Scatter plots showing the distribution of data described in Table S1. Five different single-cell RNA sequencing (scRNA-seq) cancer datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression): (a) breast cancer Dataset c, (b) breast cancer Dataset d, (c) colorectal cancer Dataset c, (d) lung cancer Dataset c, (e) ovarian cancer Dataset c. These datasets are the validation datasets. To verify if our methodology correctly classifies them according to their categories, we treated these validation datasets as unknowns even though we knew their cancer types. RNA count typically correlates with expressed genes in scRNA-seq. Deviations from this correlation, detectable via scatter plots may indicate doublets, technical errors, or low-quality cells. Insights from scatter plots inform criteria for excluding low-quality cells and adjusting data preprocessing steps.

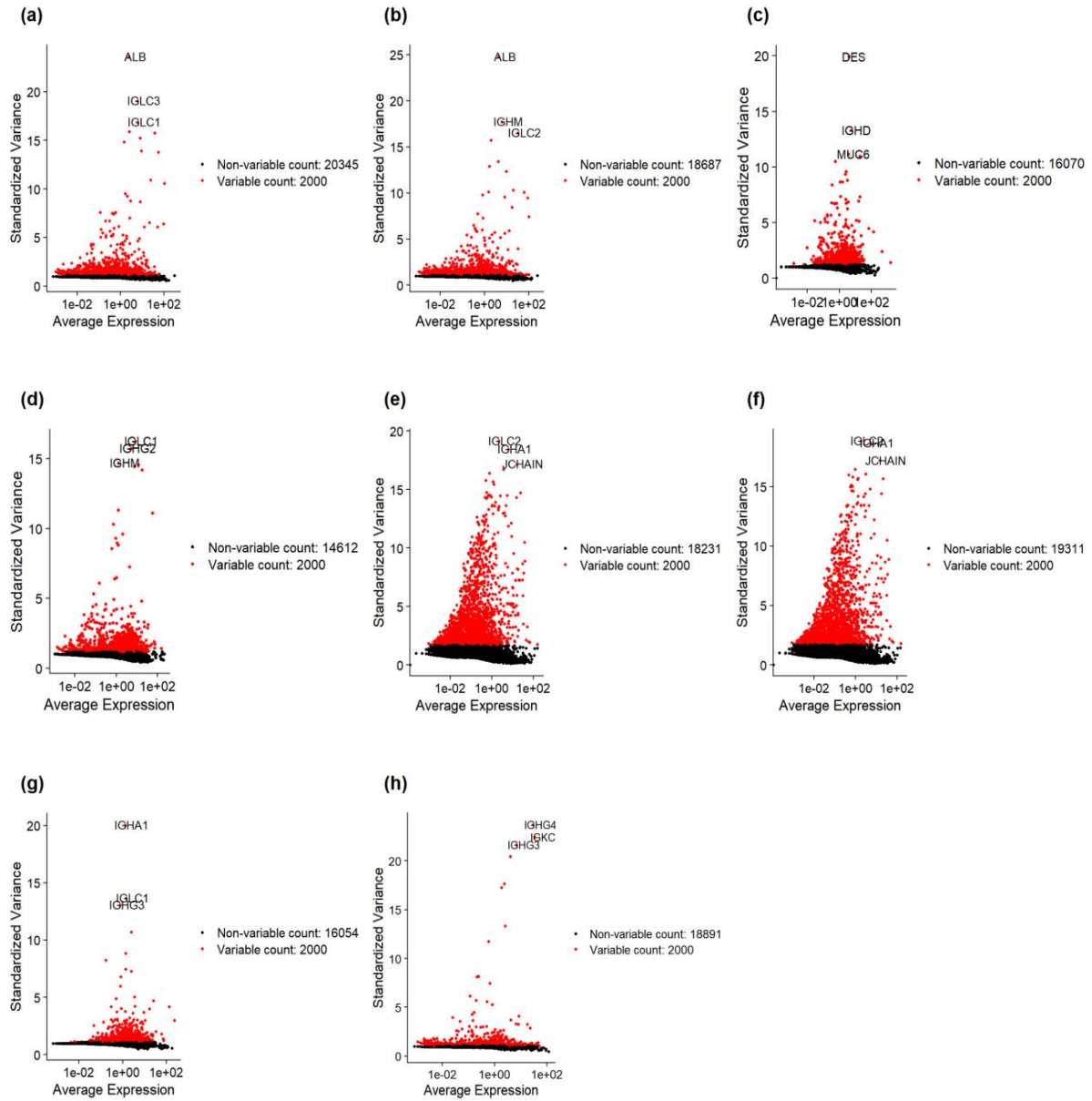


Fig. S6. Variable features plots showcasing the high variable genes in the data described in Table S1. Eight different single-cell RNA sequencing (scRNA-seq) datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression), representing four different types of cancer in Human: **(a)** breast cancer Dataset *a*, **(b)** breast cancer Dataset *b*, **(c)** colorectal cancer Dataset *a*, **(d)** colorectal cancer Dataset *b*, **(e)** lung cancer Dataset *a*, **(f)** lung cancer Dataset *b*, **(g)** ovarian cancer Dataset *a* and **(h)** ovarian cancer Dataset *b*. These datasets are the training datasets. This figure presents a visual representation of the top 2000 variable genes in the data that contribute significantly to the characterization of the respective cancer types. We applied variance stabilizing transformation (VST) method to pinpoint genes exhibiting notable expression level variations across cells post log-normalization of scRNA-seq data.

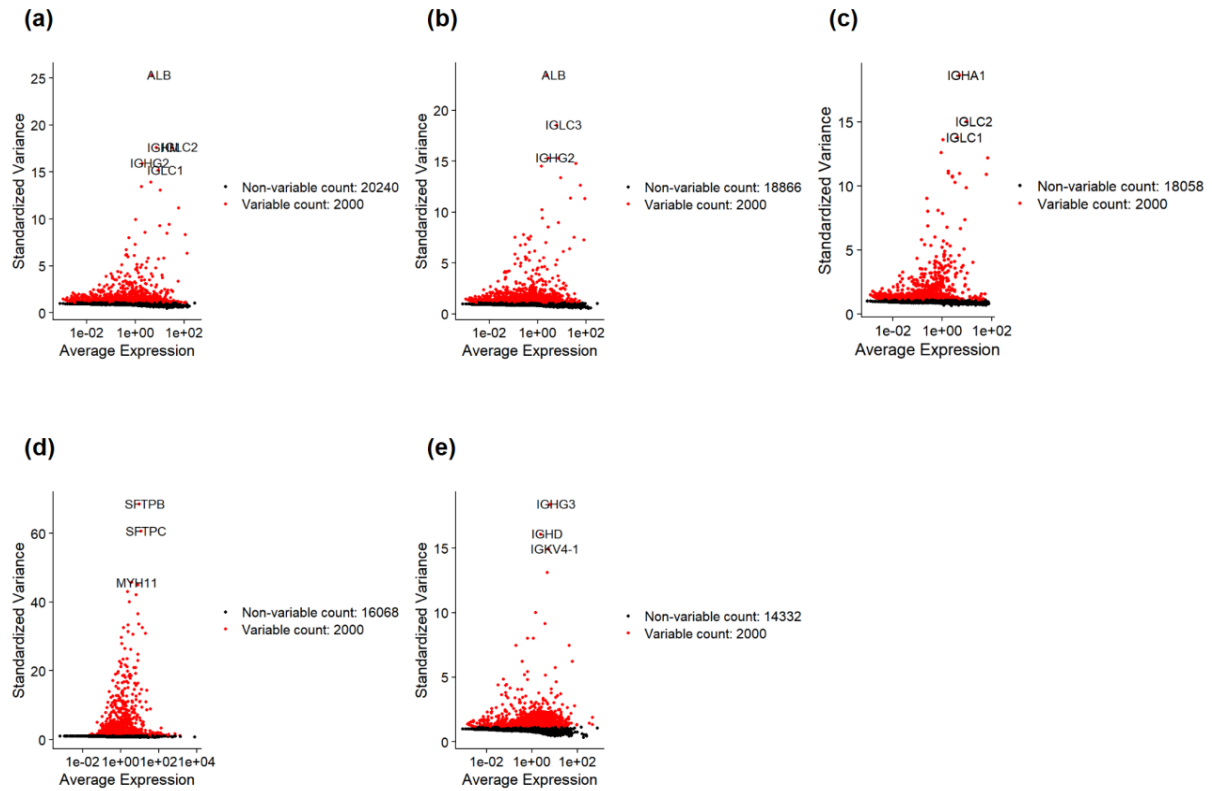


Fig. S7. Variable features plots highlighting the high variable genes in the data described in Table S1. Five different single-cell RNA sequencing (scRNA-seq) cancer datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression): (a) breast cancer Dataset c, (b) breast cancer Dataset d, (c) colorectal cancer Dataset c, (d) lung cancer Dataset c, (e) ovarian cancer Dataset c. These datasets are the validation datasets. This figure presents a visual representation of the top 2000 variable genes in the data that contribute significantly to the characterization of the respective cancer types. We applied variance stabilizing transformation (VST) method to pinpoint genes exhibiting notable expression level variations across cells post log-normalization of scRNA-seq data.

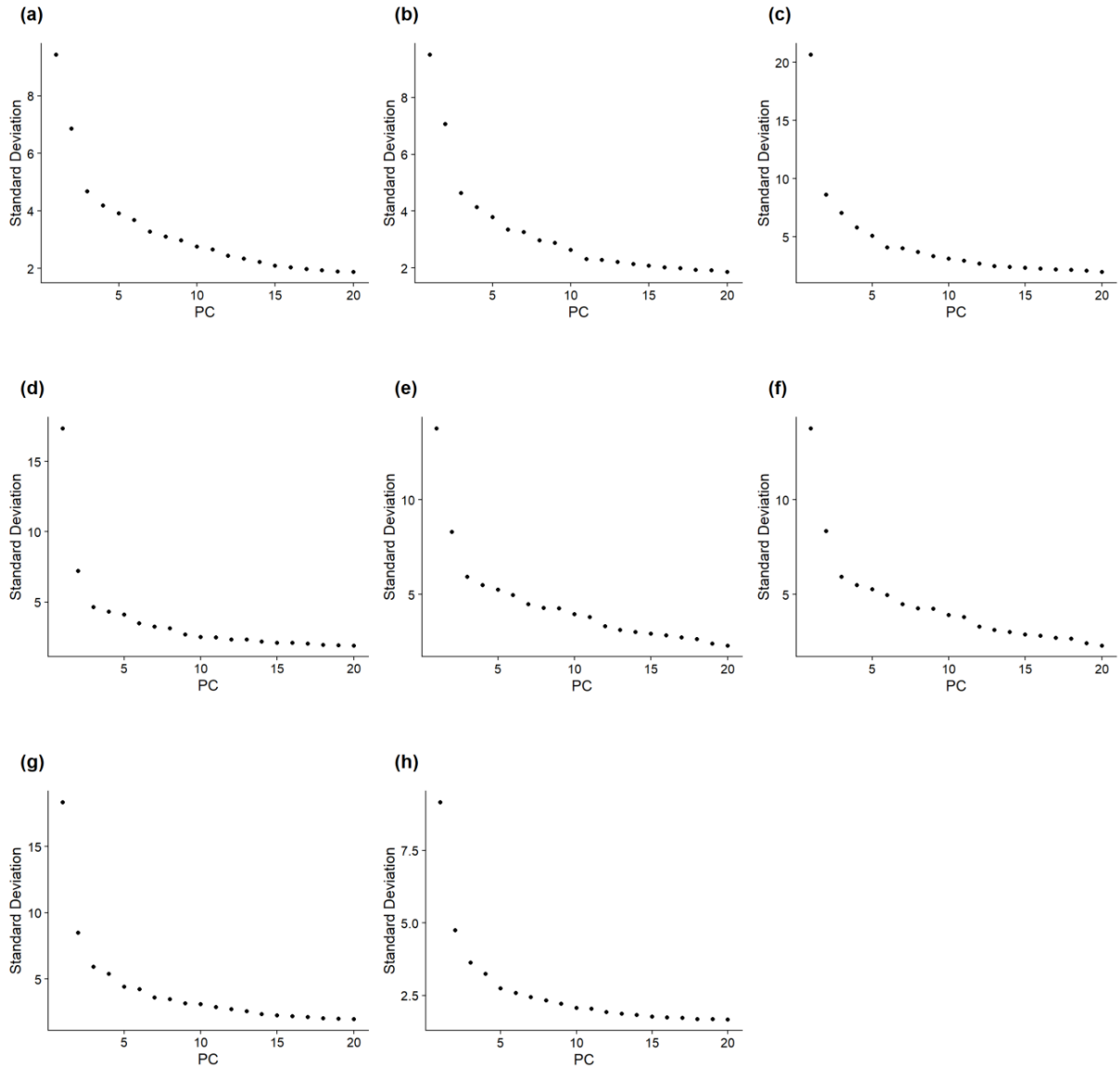


Fig. S8. Elbow plots showing Standard Deviation of PCs in the data described in Table S1. Eight different single-cell RNA sequencing (scRNA-seq) datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression), representing four different types of cancer in Human: (a) breast cancer Dataset a, (b) breast cancer Dataset b, (c) colorectal cancer Dataset a, (d) colorectal cancer Dataset b, (e) lung cancer Dataset a, (f) lung cancer Dataset b, (g) ovarian cancer Dataset a and (h) ovarian cancer Dataset b. These datasets are the training datasets. The optimal number of principal components to retain for subsequent analyses, such as clustering, visualization, or additional dimensionality reduction methods like UMAP, was determined using elbow plot. The elbow plot visually depicted the diminishing standard deviation with each additional principal component. Our focus was on identifying the “elbow” of this curve as given in Table S1, representing the point where the standard deviation decreased at a significantly slower rate.

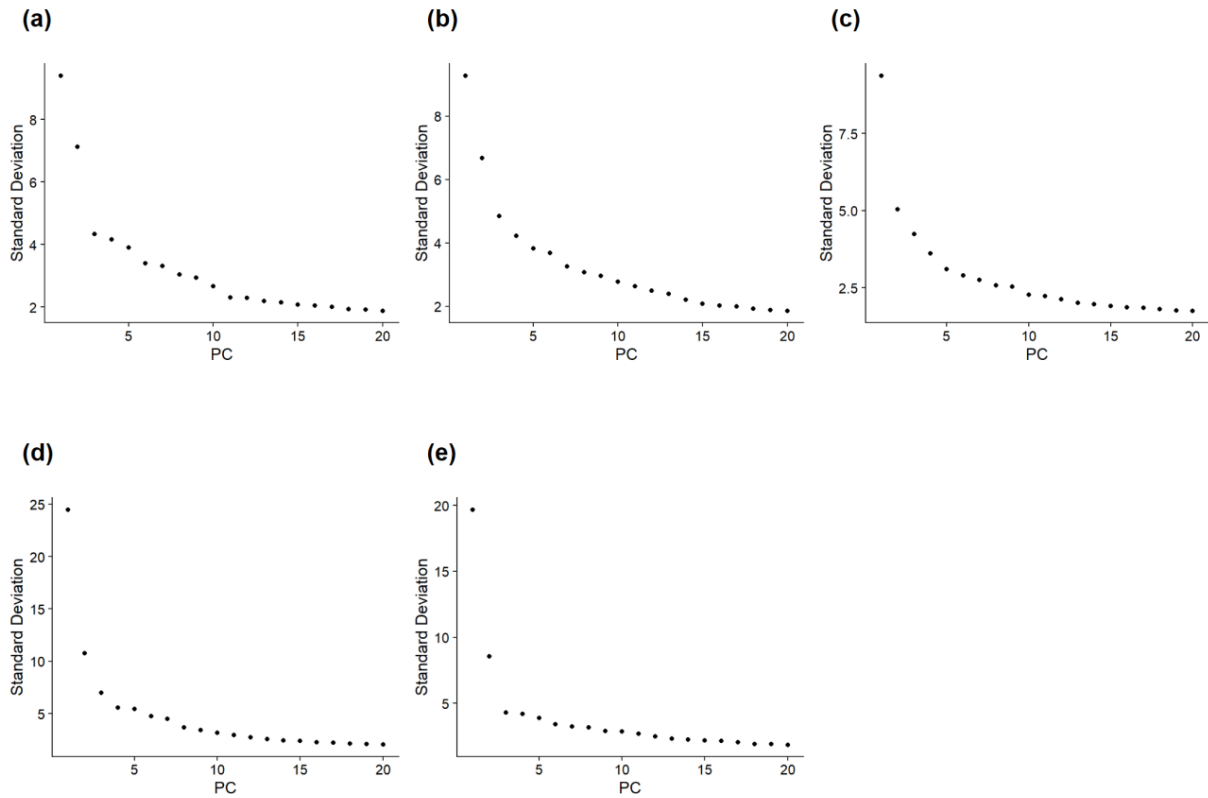


Fig. S9. Elbow plots showing Standard Deviation of PCs in the data described in Table S1. Five different single-cell RNA sequencing (scRNA-seq) datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression): (a) breast cancer Dataset c, (b) breast cancer Dataset d, (c) colorectal cancer Dataset c, (d) lung cancer Dataset c (raw), (e) ovarian cancer Dataset c. These datasets are the validation datasets. The optimal number of principal components to retain for subsequent analyses, such as clustering, visualization, or additional dimensionality reduction methods like UMAP, was determined using elbow plot. The elbow plot visually depicted the diminishing standard deviation with each additional principal component. Our focus was on identifying the “elbow” of this curve as given in Table S1, representing the point where the standard deviation decreased at a significantly slower rate.

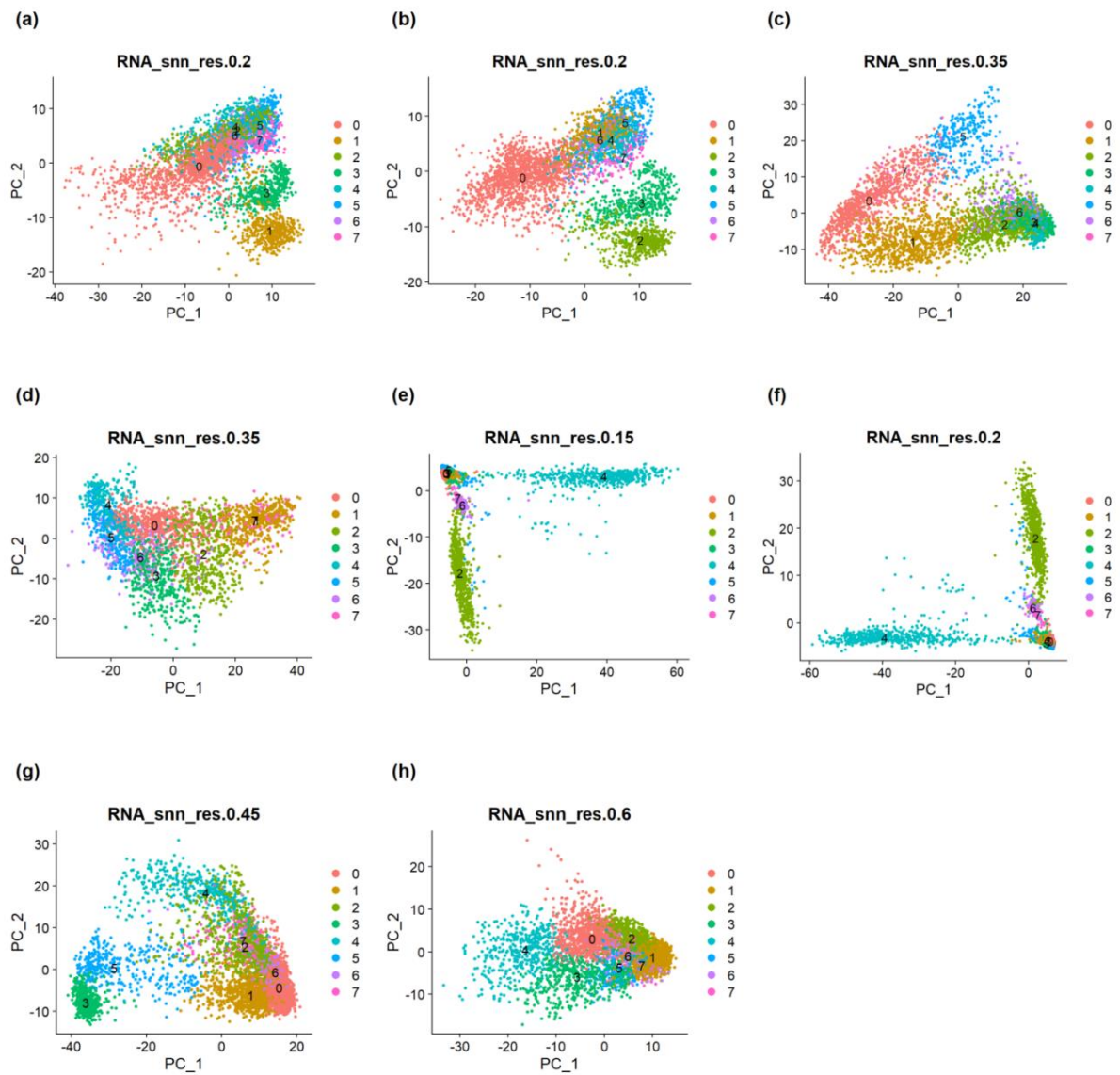


Fig. S10. Dimensional plots showing eight identified clusters derived from PCA: (a) breast cancer Dataset a, (b) breast cancer Dataset b, (c) colorectal cancer Dataset a, (d) colorectal cancer Dataset b, (e) lung cancer Dataset a, (f) lung cancer Dataset b, (g) ovarian cancer Dataset a, and (h) ovarian cancer Dataset b. These datasets are the training datasets. The elbow plots were examined to determine the number of dimensions to be utilized for the dimensional plots. Cells were grouped into distinct clusters according to gene expression patterns, aiming to identify inherent cellular subgroups or clusters within the dataset.

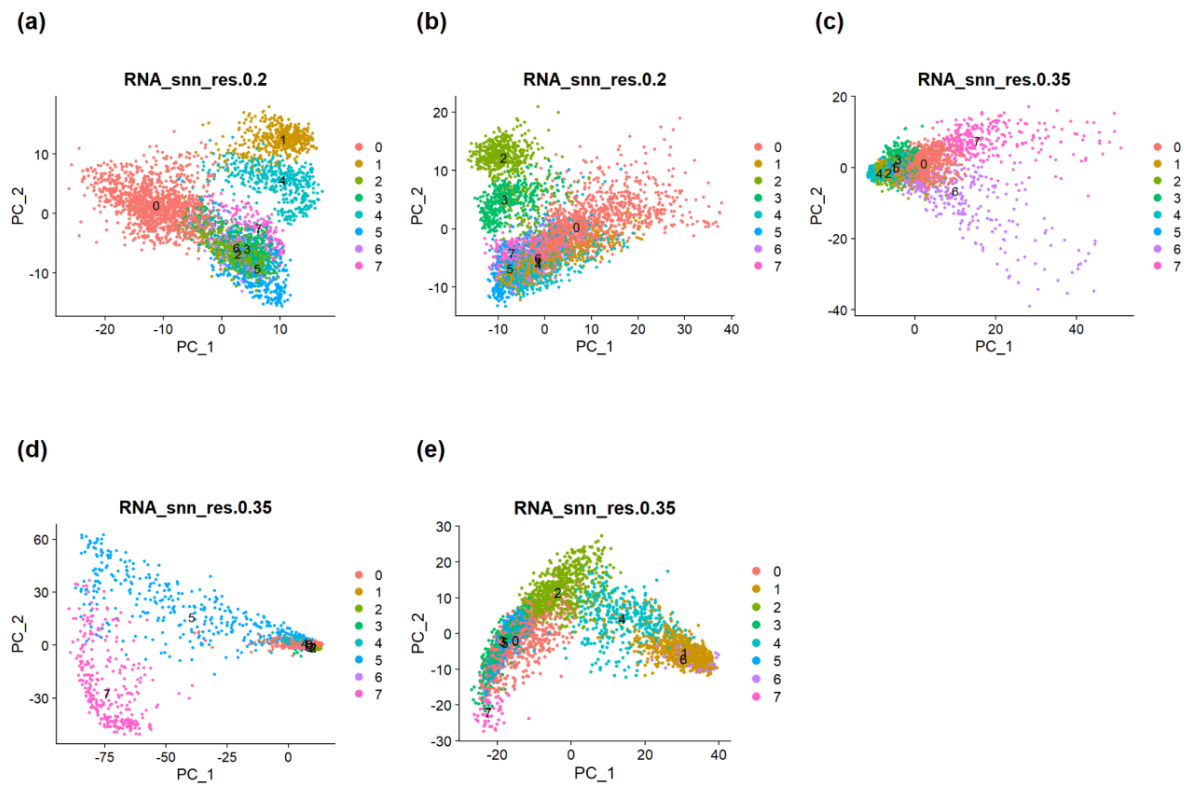


Fig. S11. Dimensional plots showing 8 identified clusters (using PCA dimension results) in the data described in Table S1. Five different single-cell RNA sequencing (scRNA-seq) datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression): (a) breast cancer Dataset c, (b) breast cancer Dataset d, (c) colorectal cancer Dataset c, (d) lung cancer Dataset c, (e) ovarian cancer Dataset c. These datasets are the validation datasets. Utilizing clustering, cells were grouped into distinct clusters according to gene expression patterns, aiming to identify inherent cellular subgroups or clusters within the dataset. We employed the Louvain algorithm as a clustering technique. The Louvain algorithm finds cohesive communities or cell clusters by optimizing the network structure.

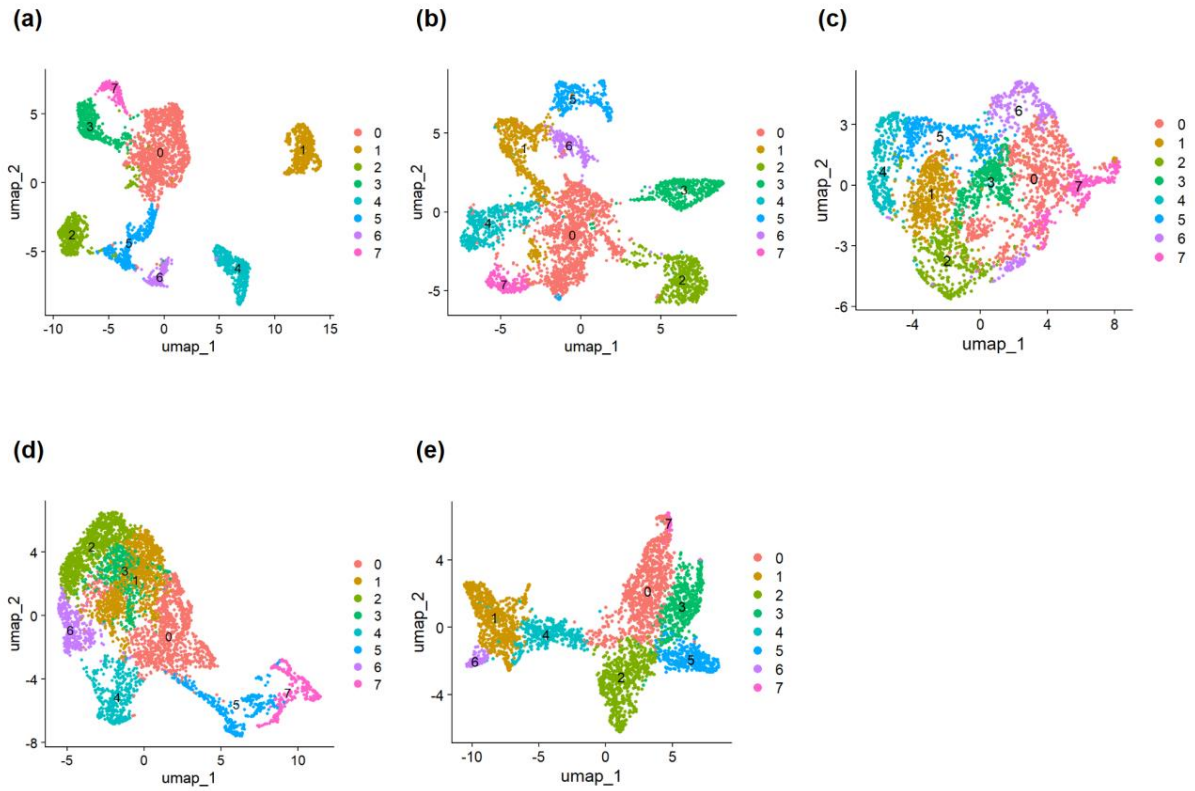


Fig. S12. UMAP Dimensional plots showing 8 identified clusters (using PCA dimension results) in the data described in Table S1. Five different single-cell RNA sequencing (scRNA-seq) datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression): (a) breast cancer Dataset c, (b) breast cancer Dataset d, (c) colorectal cancer Dataset c, (d) lung cancer Dataset c, (e) ovarian cancer Dataset c. These datasets are the validation datasets. Employing UMAP reduced dimensionality, preserving data structure visually by identifying links based on membership strengths. Visualizing with UMAP enabled cluster identification and structure observation, revealing similarities within cancer types.

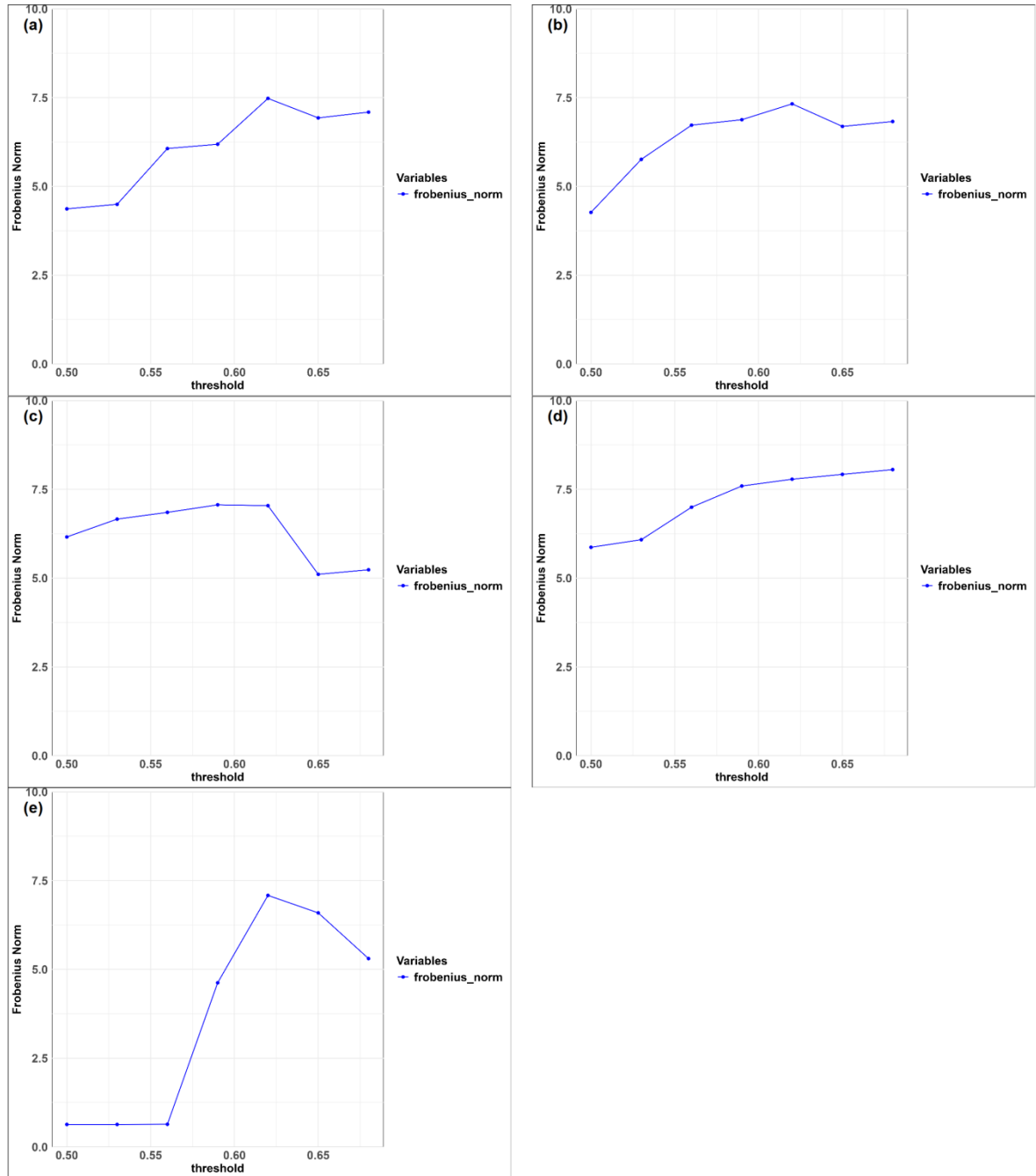


Fig. S13. Stacked line plots showing the norm patterns in the data described in Table S1. Five different single-cell RNA sequencing (scRNA-seq) datasets collated from the 10x Genomics Database (www.10xgenomics.com/products/single-cell-gene-expression): (a) breast cancer Dataset *c*, (b) breast cancer Dataset *d*, (c) colorectal cancer Dataset *c*, (d) lung cancer Dataset *c*, (e) ovarian cancer Dataset *c*. These datasets are the validation datasets. By comparing these stacked line plots and calculating the similarity percentages as given in Table S9, S10 with the training datasets (Fig. 3, Table 2), we can identify the highest similarity percentage between each random dataset and the training datasets.

Table S1. Comprehensive summary of the 8-training and 5-validation 10x genomics cancer datasets of human utilized for the cancer classification

Cancer Type	Dataset Serial Number	Disease State	Slide Serial Number	Pipeline Version	Transcriptome	Preservation Method	Number of Reads	Number of Spots Under Tissue	Mean Reads per Spot	Total Genes Detected	Median UMI Counts per Spot	Donor Count	Data Usages
breast cancer	<i>a</i>	invasive ductal carcinoma	V19L29-098-B1	Spaceranger-1.1.0	GRCh38-2020-A	Fresh Frozen	551106163	3987	138226	25072	18828	1	training
	<i>b</i>	invasive ductal carcinoma	V19L29-097-B1	Spaceranger-1.0.0	GRCh38-3.0.0	Fresh Frozen	571187331	3813	149800	22968	17531	1	training
	<i>c</i>	invasive ductal carcinoma	V19L29-097-B1	Spaceranger-1.1.0	GRCh38-2020-A	Fresh Frozen	571187331	3798	150392	24923	20762	1	validation
	<i>d</i>	invasive ductal carcinoma	V19L29-098-B1	Spaceranger-1.0.0	GRCh38-3.0.0	Fresh Frozen	551106163	4015	137262	23155	16502	1	validation
colorectal cancer	<i>a</i>	colorectal cancer	V52Y10-310-A1	Spaceranger-2.0.1	V52Y10-310-A1	FFPE*	911044482	9080	100335	18077	49282	NA*	training
	<i>b</i>	large intestine colorectal cancer	V10L13-021-B1	Spaceranger-1.3.0	GRCh38-2020-A	FFPE	205586932	2660	77288	17579	25312	NA	training
	<i>c</i>	invasive adenocarcinoma	V10A13-206-C1	Spaceranger-1.2.0	GRCh38-2020-A	Fresh Frozen	352172617	3138	112228	23233	8906	1	validation
lung cancer	<i>a</i>	non-small cell lung cancer	NA	Cellranger-2.2.0	GRCh38	Fresh	376028493	7802	48196	23148	4197	1	training
	<i>b</i>	non-small cell lung cancer	NA	Cellranger-5.0.0	GRCh38-2020-A	Fresh	376028493	7092	53022	24690	4468	1	training
	<i>c</i>	lung cancer	V52Y10-286-B1	Spaceranger-2.0.1	GRCh38-2020-A	FFPE	844866053	6195	136397	18077	66268	NA	validation
ovarian cancer	<i>a</i>	ovarian cancer	V52Y09-003-B1	Spaceranger-2.0.0	GRCh38-2020-A	FFPE	215513438	4674	46109	18064	32048	NA	training
	<i>b</i>	endometrial adenocarcinoma	V10A13-173-B1	Spaceranger-1.2.0	GRCh38-2020-A	Fresh Frozen	281487691	3493	80586	24019	8095	1	training
	<i>c</i>	ovarian cancer	V10L13-020-D1	Spaceranger-1.3.0	GRCh38-2020-A	FFPE	220554792	3455	63836	17294	28927	NA	validation

*NA: Not available

*FFPE: It is designed to measure mRNA in tissue sections derived from formalin fixed & paraffin embedded (FFPE) tissue samples and requires a Visium Spatial slide with intact tissue sections as input

**Data Source: www.10xgenomics.com/products/single-cell-gene-expression

Table S2. Comprehensive summary of the 8 training and 5 validation cancer datasets before and after filtration with quality control parameters

Cancer type	Data set Sr. No.	Sample size before filtration	Sample size after filtration	Number of features	Min features threshold*	Max features threshold*	Mitochondrial genes % threshold*	Data type	Remarks
breast	a	3986	3986	22345	NOT REQUIRED	NOT REQUIRED	NOT REQUIRED	training data	No extra filtration
	b	3813	3813	20687	NOT REQUIRED	NOT REQUIRED	NOT REQUIRED	training data	No extra filtration
	c	3798	3798	22240	NOT REQUIRED	NOT REQUIRED	NOT REQUIRED	validation data	No extra filtration
	d	4014	4014	20866	NOT REQUIRED	NOT REQUIRED	NOT REQUIRED	validation data	No extra filtration
colorectal	a	9079	4067	18070	1000	9500	10	training data	Subset of the filtered data
	b	2660	2660	16612	NOT REQUIRED	NOT REQUIRED	NOT REQUIRED	training data	No extra filtration
	c	3138	3138	20058	NOT REQUIRED	NOT REQUIRED	NOT REQUIRED	validation data	No extra filtration
lung	a	7591	4756	20231	1100	7500	6	training data	Subset of the filtered data
	b	6949	4803	21311	1100	7500	6	training data	Subset of the filtered data
	c	6195	4760	18068	1000	11500	8	validation data	Subset of the filtered data
ovarian	a	4671	4671	18054	NOT REQUIRED	NOT REQUIRED	NOT REQUIRED	training data	No extra filtration
	b	3491	3491	20891	NOT REQUIRED	NOT REQUIRED	NOT REQUIRED	training data	No extra filtration
	c	3455	3455	16332	NOT REQUIRED	NOT REQUIRED	NOT REQUIRED	validation data	No extra filtration

* To establish a baseline for further comparative analysis by trail-and error method, min-and max feature thresholds and Mitochondrial genes% threshold were defined to keep the filtered sample size for all the considered datasets within the range of 2500-5000

Data source: www.10xgenomics.com/products/single-cell-gene-expression

Table S3. The number of PCs in the PCA taken by observing elbow plots and the resolution parameters adjusted to get 8 clusters of the 8 training and 5 validation cancer datasets

Cancer type	Dataset Serial No.	Dimension (Number of PCs)	Resolution Parameter	Datatype
breast	<i>a</i>	14	0.20	training data
	<i>b</i>	10	0.20	training data
	<i>c</i>	12	0.20	validation data
	<i>d</i>	14	0.20	validation data
colorectal	<i>a</i>	12	0.35	training data
	<i>b</i>	13	0.35	training data
	<i>c</i>	12	0.35	validation data
lung	<i>a</i>	18	0.15	training data
	<i>b</i>	18	0.20	training data
	<i>c</i>	12	.35	validation data
ovarian	<i>a</i>	13	0.45	training data
	<i>b</i>	12	0.60	training data
	<i>c</i>	13	0.35	validation data

Table S4. The number of cells present in different clusters of the 8 training and 5 validation cancer datasets with the Mean Silhouette Score of the clusters

Cancer type	Dataset No.	Cluster								Mean Silhouette Score	Datatype
		0	1	2	3	4	5	6	7		
breast	<i>a</i>	1300	524	521	458	445	333	207	198	0.472	training data
	<i>b</i>	1318	518	492	430	404	316	180	155	0.532	training data
	<i>c</i>	1280	478	469	428	425	367	182	169	0.567	validation data
	<i>d</i>	1291	556	538	458	441	329	202	199	0.404	validation data
colorectal	<i>a</i>	905	851	844	507	338	297	284	41	0.293	training data
	<i>b</i>	568	447	407	393	289	250	216	90	0.221	training data
	<i>c</i>	742	466	386	355	320	317	279	273	0.187	validation data
lung	<i>a</i>	1737	908	728	563	527	167	89	37	0.647	training data
	<i>b</i>	1750	902	733	583	534	173	90	38	0.679	training data
	<i>c</i>	1070	1001	868	433	406	363	324	295	0.236	validation data
ovarian	<i>a</i>	1293	783	766	597	422	362	283	165	0.343	training data
	<i>b</i>	914	750	525	475	428	203	174	22	0.194	training data
	<i>c</i>	788	708	643	472	375	341	82	46	0.333	validation data

Table S5. The number of genes present in the 8 distinct networks of the training and validation cancer datasets at gene correlation threshold 0.6 and Shannon index of genes in the networks

Cancer type	Dataset No.	Cluster								Shannon index	Datatype
		0	1	2	3	4	5	6	7		
breast	<i>a</i>	31	49	92	73	100	57	47	22	2.857	training data
	<i>b</i>	14	47	89	113	58	124	49	45	2.793	training data
	<i>c</i>	21	89	67	67	121	85	54	47	2.871	validation data
	<i>d</i>	35	72	48	69	97	55	45	20	2.877	validation data
colorectal	<i>a</i>	266	292	222	81	147	114	142	757	2.633	training data
	<i>b</i>	76	153	148	79	75	109	25	329	2.691	training data
	<i>c</i>	25	25	33	20	28	22	95	97	2.685	validation data
lung	<i>a</i>	15	93	177	44	420	1297	100	569	2.135	training data
	<i>b</i>	12	100	167	42	429	1268	101	533	2.138	training data
	<i>c</i>	44	13	14	7	154	335	20	120	2.103	validation data
ovarian	<i>a</i>	70	67	85	12	91	16	69	101	2.787	training data
	<i>b</i>	22	23	23	32	29	34	26	1621	0.774	training data
	<i>c</i>	128	52	55	81	62	63	56	315	2.634	validation data

Table S6. The number of edges present in the 8 distinct networks of the training and validation cancer datasets at gene correlation threshold 0.6 and Shannon index of edges in the networks

Cancer type	Dataset No.	Cluster								Shannon index	Datatype
		0	1	2	3	4	5	6	7		
breast	<i>a</i>	51	85	182	104	159	81	74	53	2.860	training data
	<i>b</i>	11	66	113	204	89	180	68	59	2.716	training data
	<i>c</i>	16	124	111	97	233	121	72	65	2.769	validation data
	<i>d</i>	53	121	82	100	145	78	75	54	2.920	validation data
colorectal	<i>a</i>	1659	1344	574	144	422	442	478	3873	2.384	training data
	<i>b</i>	135	516	269	168	117	242	30	3026	1.727	training data
	<i>c</i>	46	43	57	37	45	42	154	213	2.639	validation data
lung	<i>a</i>	14	327	558	50	2980	48569	154	891	0.608	training data
	<i>b</i>	9	382	515	52	3009	46963	162	982	0.636	training data
	<i>c</i>	41	9	12	6	581	4549	13	287	0.910	validation data
ovarian	<i>a</i>	145	163	188	8	111	22	123	252	2.671	training data
	<i>b</i>	30	45	51	54	48	53	62	3157	0.735	training data
	<i>c</i>	206	106	63	98	98	66	81	461	2.603	validation data

Table S7. Table of the cluster no. 0 network of the breast cancer Dataset *b* at gene correlation threshold 0.6

Gene column 1	Gene column 2	Correlation value
IGHG4	IGKC	0.707
IGHG1	IGKC	0.746
IGHG3	IGKC	0.757
KRT19	KRT8	0.633
IGHG1	IGHG4	0.675
IGHG3	IGHG4	0.661
IGHG3	IGHG1	0.708
IGLC2	IGHG1	0.634
IGLC2	IGHG3	0.698
MT-CO3	MT-CO2	0.612
MT-CYB	MT-CO3	0.613

Table S8. The norms of the Hausdorff distance matrices of the validation cancer datasets at 6 gene correlation thresholds

Threshold	Norm B(c)	Norm B(d)	Norm C(c)	Norm L(c)	Norm O(c)
0.5	4.369	4.270	6.160	5.869	0.637
0.53	4.497	5.763	6.662	6.082	0.637
0.56	6.068	6.725	6.851	6.998	0.640
0.59	6.189	6.879	7.064	7.594	4.625
0.62	7.479	7.325	7.040	7.785	7.084
0.65	6.930	6.690	5.108	7.922	6.592
0.68	7.092	6.828	5.236	8.056	5.305
B: breast; L: lung; C: colorectal; O: ovarian; (c),(d): Dataset id.					

Table S9. Interval wise and overall normalized similarity percentages results of the successful test (validation) datasets with training datasets

% Similarity	0.50-0.53	0.53-0.56	0.56-0.59	0.59-0.62	0.62-0.65	0.65-0.68	Overall
breast(a),Testdata1	86.55	84.76	89.53	92.36	98.90	98.61	91.79
breast(b),Testdata1	85.93	83.92	85.84	88.16	98.98	98.86	90.28
colorectal(a),Testdata1	49.70	36.57	25.48	41.56	76.02	87.83	52.86
colorectal(b),Testdata1	46.75	33.68	22.54	44.57	94.65	92.30	55.75
lung(a),Testdata1	78.28	82.04	83.52	86.12	91.05	87.93	84.82
lung(b),Testdata1	97.83	90.43	83.13	85.81	90.67	87.15	89.17
ovarian(a),Testdata1	78.60	77.53	92.79	76.85	71.85	76.69	79.05
ovarian(b),Testdata1	74.31	76.66	87.45	88.08	93.84	99.28	86.60
breast(a),Testdata2	99.16	99.15	99.17	98.84	98.22	97.79	98.72
breast(b),Testdata2	98.30	98.28	95.02	94.72	96.11	95.27	96.28
colorectal(a),Testdata2	40.41	23.53	15.85	35.69	77.52	85.09	46.35
colorectal(b),Testdata2	37.51	20.60	12.91	37.64	92.76	88.71	48.35
lung(a),Testdata2	83.48	96.31	92.84	92.19	88.26	84.33	89.57
lung(b),Testdata2	88.31	88.27	92.47	91.89	87.88	83.55	88.73
ovarian(a),Testdata2	91.41	91.47	93.99	77.61	74.48	80.29	84.88
ovarian(b),Testdata2	83.16	90.59	96.69	94.38	94.55	95.68	92.51
breast(a),Testdata4	84.16	96.61	93.06	92.46	88.45	84.64	89.90
breast(b),Testdata4	84.83	97.48	97.14	96.52	90.49	87.17	92.27
colorectal(a),Testdata4	27.65	19.34	8.71	27.28	70.89	89.00	40.48
colorectal(b),Testdata4	24.71	16.41	5.77	30.20	92.37	93.58	43.84
lung(a),Testdata4	99.39	99.38	99.43	99.20	98.46	98.10	98.99
lung(b),Testdata4	80.11	85.39	99.75	99.43	98.83	98.82	93.72
ovarian(a),Testdata4	81.97	94.25	86.60	70.63	63.02	62.72	76.53
ovarian(b),Testdata4	93.47	93.49	95.50	89.94	86.75	86.75	90.98

Table S10. Interval wise and overall normalized similarity percentages results of the failed test (validation) datasets with training datasets

% Similarity	0.50-0.53	0.53-0.56	0.56-0.59	0.59-0.62	0.62-0.65	0.65-0.68	Overall
breast(a),Testdata3	79.61	91.55	98.57	95.97	82.11	75.12	87.15
breast(b),Testdata3	80.40	92.21	97.01	94.20	80.10	72.59	86.09
lung(a),Testdata3	21.37	16.50	13.63	34.86	73.64	64.40	37.40
lung(b),Testdata3	18.43	13.56	10.69	35.75	76.74	66.06	36.87
colorectal(a),Testdata3	92.98	93.42	95.02	91.56	72.13	61.66	84.46
colorectal(b),Testdata3	73.75	79.48	94.65	91.32	71.75	60.87	78.64
ovarian(a),Testdata3	79.82	99.04	92.11	79.72	81.15	97.01	88.14
ovarian(b),Testdata3	94.88	99.30	98.48	97.80	81.00	73.01	90.75
breast(a),Testdata5	35.69	18.79	33.72	76.34	95.79	82.80	57.19
breast(b),Testdata5	34.84	17.92	31.56	72.01	93.77	80.72	55.14
lung(a),Testdata5	96.06	96.09	62.40	57.51	79.83	69.68	76.93
lung(b),Testdata5	99.01	99.03	60.41	60.39	91.67	74.56	80.84
colorectal(a),Testdata5	24.03	15.96	29.50	69.84	86.35	70.88	49.43
colorectal(b),Testdata5	43.60	24.07	29.15	69.53	85.99	70.06	53.73
ovarian(a),Testdata5	29.97	13.54	35.80	74.53	77.06	87.32	53.04
ovarian(b),Testdata5	20.56	12.82	32.66	73.89	95.51	81.04	52.75

Table S11. Overall normalized similarity percentages results between similar and different cancer type datasets including both training and validation datasets

Dataset 1	Dataset 2	Overall % Similarity
breast(a)	breast(b)	97.5
breast(a)	colorectal(a)	45.91
breast(a)	colorectal(b)	48.4
breast(a)	lung(a)	90.77
breast(a)	lung(b)	89.51
breast(a)	ovarian(a)	84.15
breast(a)	ovarian(b)	93.11
breast(a)	breast(c)	91.79
breast(a)	breast(d)	98.72
breast(a)	colorectal(c)	87.15
breast(a)	lung(c)	89.89
breast(a)	ovarian(c)	57.19
breast(b)	colorectal(a)	44.4
breast(b)	colorectal(b)	47.42
breast(b)	lung(a)	93.14
breast(b)	lung(b)	91.45
breast(b)	ovarian(a)	82.46
breast(b)	ovarian(b)	93.5
breast(b)	breast(c)	90.28
breast(b)	breast(d)	96.28
breast(b)	colorectal(c)	86.09
breast(b)	lung(c)	92.27
breast(b)	ovarian(c)	55.14
colorectal(a)	colorectal(b)	89.85
colorectal(a)	lung(a)	41.16
colorectal(a)	lung(b)	45.59
colorectal(a)	ovarian(a)	44.73
colorectal(a)	ovarian(b)	43.12
colorectal(a)	breast(c)	52.86
colorectal(a)	breast(d)	46.35
colorectal(a)	colorectal(c)	37.4
colorectal(a)	lung(c)	40.48
colorectal(a)	ovarian(c)	76.93
colorectal(b)	lung(a)	44.71
colorectal(b)	lung(b)	49.03
colorectal(b)	ovarian(a)	40.3
colorectal(b)	ovarian(b)	44.52
colorectal(b)	breast(c)	55.75
colorectal(b)	breast(d)	48.35
colorectal(b)	colorectal(c)	36.87
colorectal(b)	lung(c)	43.84
colorectal(b)	ovarian(c)	80.84
lung(a)	lung(b)	94.08
lung(a)	ovarian(a)	77.17
lung(a)	ovarian(b)	91.56
lung(a)	breast(c)	84.82
lung(a)	breast(d)	89.57
lung(a)	colorectal(c)	84.46
lung(a)	lung(c)	98.99
lung(a)	ovarian(c)	49.43
lung(b)	ovarian(a)	74.28
lung(b)	ovarian(b)	86.23
lung(b)	breast(c)	89.17
lung(b)	breast(d)	88.73

Dataset 1	Dataset 2	Overall % Similarity
lung(b)	colorectal(c)	78.64
lung(b)	lung(c)	93.72
lung(b)	ovarian(c)	53.73
ovarian(a)	ovarian(b)	84.62
ovarian(a)	breast(c)	79.05
ovarian(a)	breast(d)	84.87
ovarian(a)	colorectal(c)	88.14
ovarian(a)	lung(c)	76.53
ovarian(a)	ovarian(c)	53.04
ovarian(b)	breast(c)	86.6
ovarian(b)	breast(d)	92.51
ovarian(b)	colorectal(c)	90.75
ovarian(b)	lung(c)	90.98
ovarian(b)	ovarian(c)	52.75
breast(c)	breast(d)	91.58
breast(c)	colorectal(c)	80.24
breast(c)	lung(c)	83.9
breast(c)	ovarian(c)	63.58
breast(d)	colorectal(c)	87.52
breast(d)	lung(c)	88.69
breast(d)	ovarian(c)	58.28
colorectal(c)	lung(c)	83.83
colorectal(c)	ovarian(c)	51.33
lung(c)	ovarian(c)	48.57

9. USER GUIDE FOR “analyze10xgenomCancer” FUNCTION

Table S12. List of all user defined parameters for the computational analysis of 10x genomics dataset

Parameter	Values	Parameter type	Remark
min.cells	3	Default	Seurat object formation
min.features	100	Default	Seurat object formation
Min features threshold	*	User defined	Filtration parameter
Max features threshold	*	User defined	Filtration
Mitochondrial genes % threshold	*	User defined	Filtration
Scale factor	10000	Default	Data Normalization
nfeatures	2000	Default	High variable features identification
dimension	*	User defined	Number of PCs
Resolution Parameter	*	User defined	Clustering
alpha	0.05	Default	Correlation network parameter
Gene correlation threshold	0.50-0.68	Default	One at a time with the increment of 0.03
Max possible RMSE value	7	Default	Outcome normalization parameter
Note: The R function “analyze10xgenomCancer()” accepts user defined inputs (*) for these parameters; all other parameters use default values			

Function Execution Example

For a filtered dataset within the sample size range 2500-5000 obtained from the 10x Genomics Database, set the `min_features=0`, `max_features= 25000`, `mt_threshold= 100` in the following R-function after running the function to use the whole dataset information. The 'dims' and 'resolution' are user-defined parameters. In this example, we chose a breast cancer dataset, which is of size 3,813; so we set the parameters as mentioned in Table S12. We set `dims=1:10` and `resolution=0.2` for this particular example. In this case, the following command will be fine:

```
Results = analyze10xgenomCancer(file\_path, 0, 25000, 100, 1:10, 0.2)
```

For a filtered dataset with a sample size larger than 5000 obtained from the 10x Genomics Database, the `min_features`, `max_features`, `mt_threshold`, `dims`, and `resolution` are all user-defined parameters. In this case, the following command will be ok with setting the parameters by the user (see Table S2, S3 for details).

```
Results= analyze10xgenomCancer(file\_path, min\_features, max\_features,  
mt\_threshold, dims, resolution)
```

The above function will return a UMAP plot, a line plot, and finally a **Frobenius norms table** at different gene correlation thresholds. If something wrong happens, the error information will be displayed.

Overview

The ***analyze10xgenomCancer*** function in R is designed to process 10x Genomics single-cell RNA-seq data, specifically aimed at cancer research. This function handles data loading, preprocessing, clustering, and visualization, along with advanced analyses like gene correlation network construction and the calculation of Hausdorff distances for cluster comparisons, and Cancer classification.

Usage

The function can be used in a default mode with minimal required inputs but also allows for extensive customization through various parameters.

Parameters

- **file_path**: Path to the file containing the gene expression matrix (HDF5 format).
- **min_features** (default = 0): Minimum number of features (genes) to filter cells.
- **max_features** (default = 25000): Maximum number of features (genes) for cell filtering.
- **mt_threshold** (default = 100): Threshold for mitochondrial content filtering. Cells above this threshold will be excluded.
- **dims** (default = 1:10): Dimensions to use for PCA and UMAP.
- **resolution** (default = 0.2): Clustering resolution for identifying cell populations.

Example Usage

Here are two example commands to run the ***analyze10xgenomCancer*** function based on different user needs:

1. **Basic Usage** (no advanced configurations, just data processing):

```
analyze10xgenomCancer("file_path  
\\V1_Breast_Cancer_Block_A_Section_1_filtered_feature_bc_matrix.h5")
```

2. **Advanced Usage** (with custom configurations for detailed analysis):

```
analyze10xgenomCancer(          file_path          =          "file_path  
\\V1_Breast_Cancer_Block_A_Section_1_filtered_feature_bc_matrix.h5",  min_features  =  
200, max_features = 6000, mt_threshold = 10, dims = 1:15, resolution = 0.3 )
```

Output

- **SeuratObject**: A Seurat object containing the processed single-cell data.
- **UMAPPlot**: A plot object showing the UMAP visualization of the data.

- **Hausdorff distance matrices:** Distance matrix in different threshold
- **Distribution of Frobenius Norm:** Changes of Frobenius Norm in different threshold.
- **LinePlot:** A plot object showing changes of Frobenius Norm in different threshold.

Error Handling

If there is an issue with the data input or during any of the processing steps, the function will return an error message detailing what went wrong, ensuring that users can make necessary adjustments.

REFERENCES

1. Ordoñez-Rueda D., Baying B., Pavlinic D., Alessandri L., Yeboah Y., Landry J.J., Calogero R., Benes V., Paulsen M. Apoptotic cell exclusion and bias-free single-cell selection are important quality control requirements for successful single-cell sequencing applications. *Cytom. A*. 2020. V. 97. No. 2. P. 156–167. doi: [10.1002/cyto.a.23898](https://doi.org/10.1002/cyto.a.23898)
2. Stuart T., Butler A., Hoffman P., Hafemeister C., Papalexi E., Mauck W.M., Hao Y., Stoeckius M., Smibert P., Satija R. Comprehensive integration of single-cell data. *Cell*. 2019. V. 177. No. 7. P. 1888–1902. doi: [10.1016/j.cell.2019.05.031](https://doi.org/10.1016/j.cell.2019.05.031)
3. Salazar-Martín A.G., Kalluri A.S., Villanueva M.A., Hughes T.K., Wadsworth M.H., Dao T.T., Balcells M., Nezami F.R., Shalek A.K., Edelman E.R. Single-cell RNA sequencing reveals that adaptation of human aortic endothelial cells to antiproliferative therapies is modulated by flow-induced shear stress. *Arterioscler. Thromb. Vasc. Biol.* 2023. V. 43. No. 12. P. 2265–2281. doi: [10.1161/ATVBAHA.123.319283](https://doi.org/10.1161/ATVBAHA.123.319283)
4. Andrews T.S., Hemberg M. Identifying cell populations with scRNASeq. *Mol. Aspects Med.* 2018. V. 59. P. 114–122. doi: [10.1016/j.mam.2017.07.002](https://doi.org/10.1016/j.mam.2017.07.002)
5. Schmidt F., Ranjan B., Lin Q.X., Krishnan V., Joanito I., Honardoost M.A., Nawaz Z., Venkatesh P.N., Tan J., Rayan N.A., Ong S.T. RCA2: a scalable supervised clustering algorithm that reduces batch effects in scRNA-seq data. *Nucleic Acids Res.* 2021. V. 49. No. 15. P. 8505–8519. doi: [10.1093/nar/gkab632](https://doi.org/10.1093/nar/gkab632)
6. Grabski I.N., Street K., Irizarry R.A. Significance analysis for clustering with single-cell RNA-sequencing data. *Nat. Methods.* 2023. V. 20. No. 8. P. 1196–1202. doi: [10.1038/s41592-023-01933-9](https://doi.org/10.1038/s41592-023-01933-9)