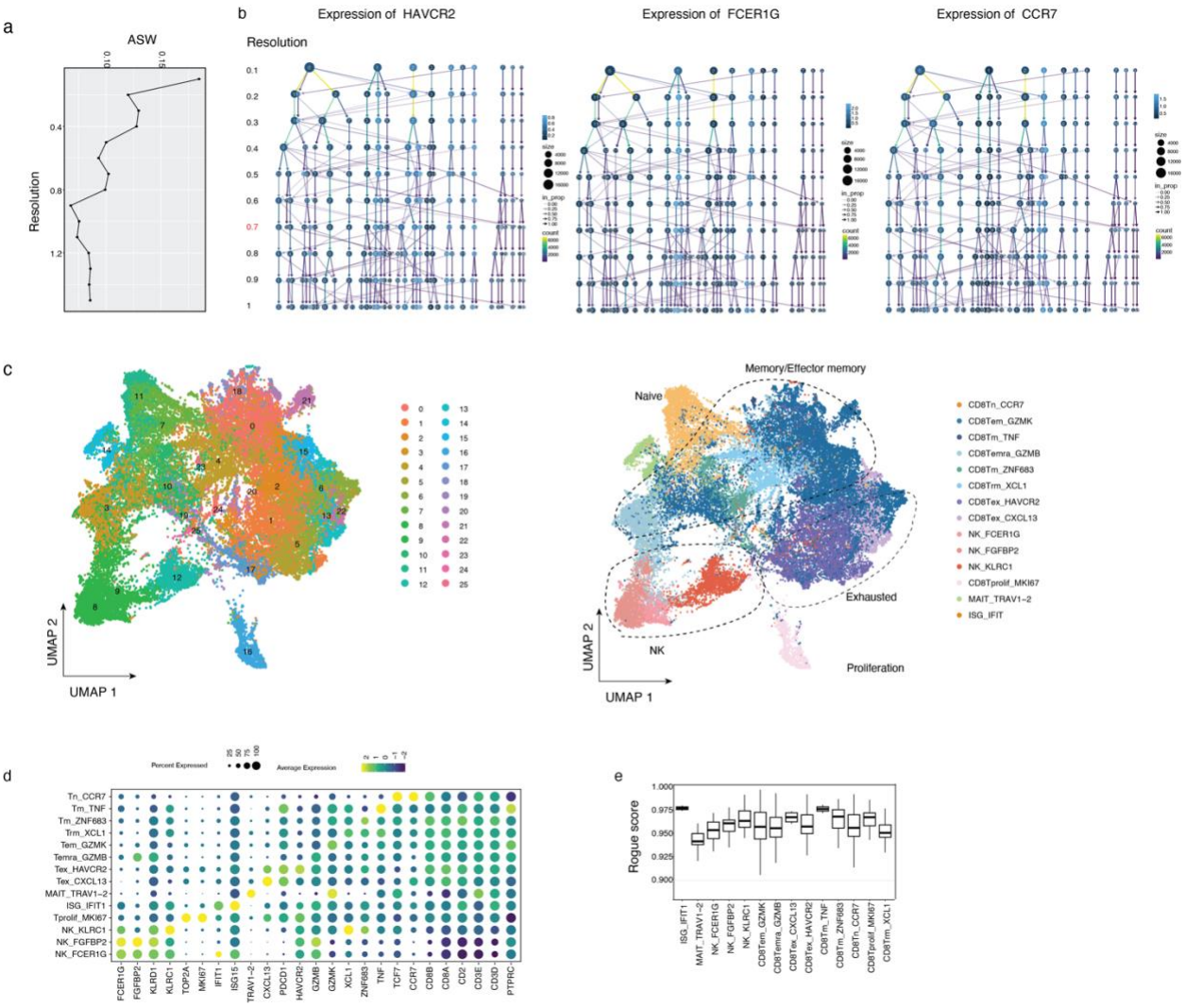
Supplementary Materials

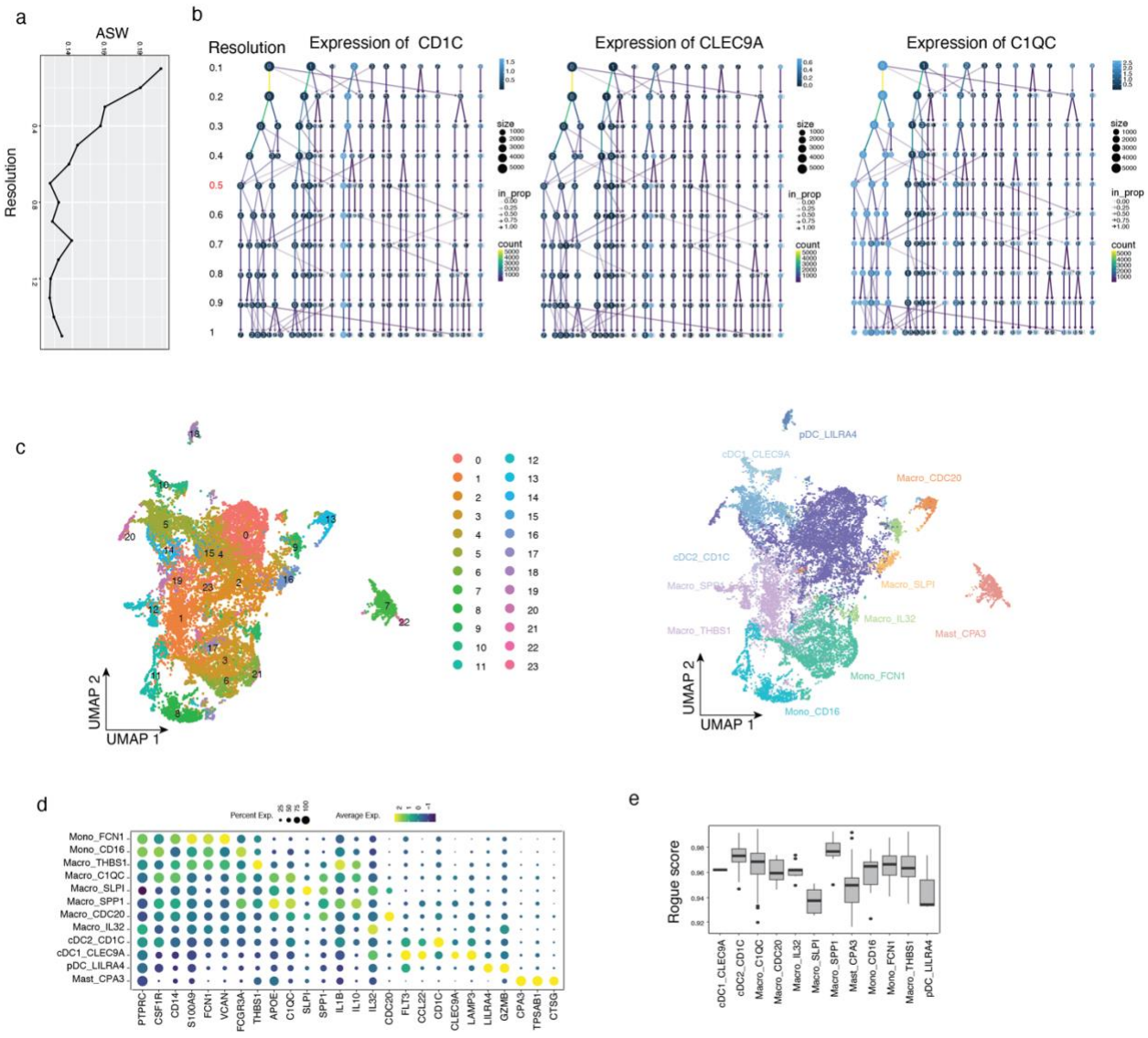
Comprehensive evaluation of optimal cluster resolution across distinct cell lineages

To address the risk of over-clustering, we employed Silhouette scores and Clustree metrics to ascertain the optimal resolution for clustering. To mitigate the curse of dimensionality, we derived distances from the principal component analysis (PCA) space representation. We then evaluated the variation of the average silhouette width using Euclidean distance metrics, considering resolutions ranging from 0.1 to 1.5 in each cell lineage. These measurements were derived from the Euclidean distances calculated within the top 30 principal components of the PCA space. We deliberately chose the number of principal components to maintain consistency with the actual number used in the clustering and visualization processes. After obtaining the best resolution from the silhouette width for specific cell lineages, we observed that the expression of several classic cell type marker genes was co-mingled with other cell types. If the expression of classic marker genes exhibited distinct upregulation in separate clusters at the optimal resolution identified by average silhouette width (ASW), it indicated that the current resolution was suitable for distinguishing the specific lineage. Conversely, if marker gene expressions were intermixed with markers of other cell types, relying solely on ASW would be inadequate for effectively delineating subcellular populations. Therefore, we integrated Clustree analysis with marker gene expression to determine the most appropriate resolution for each lineage.

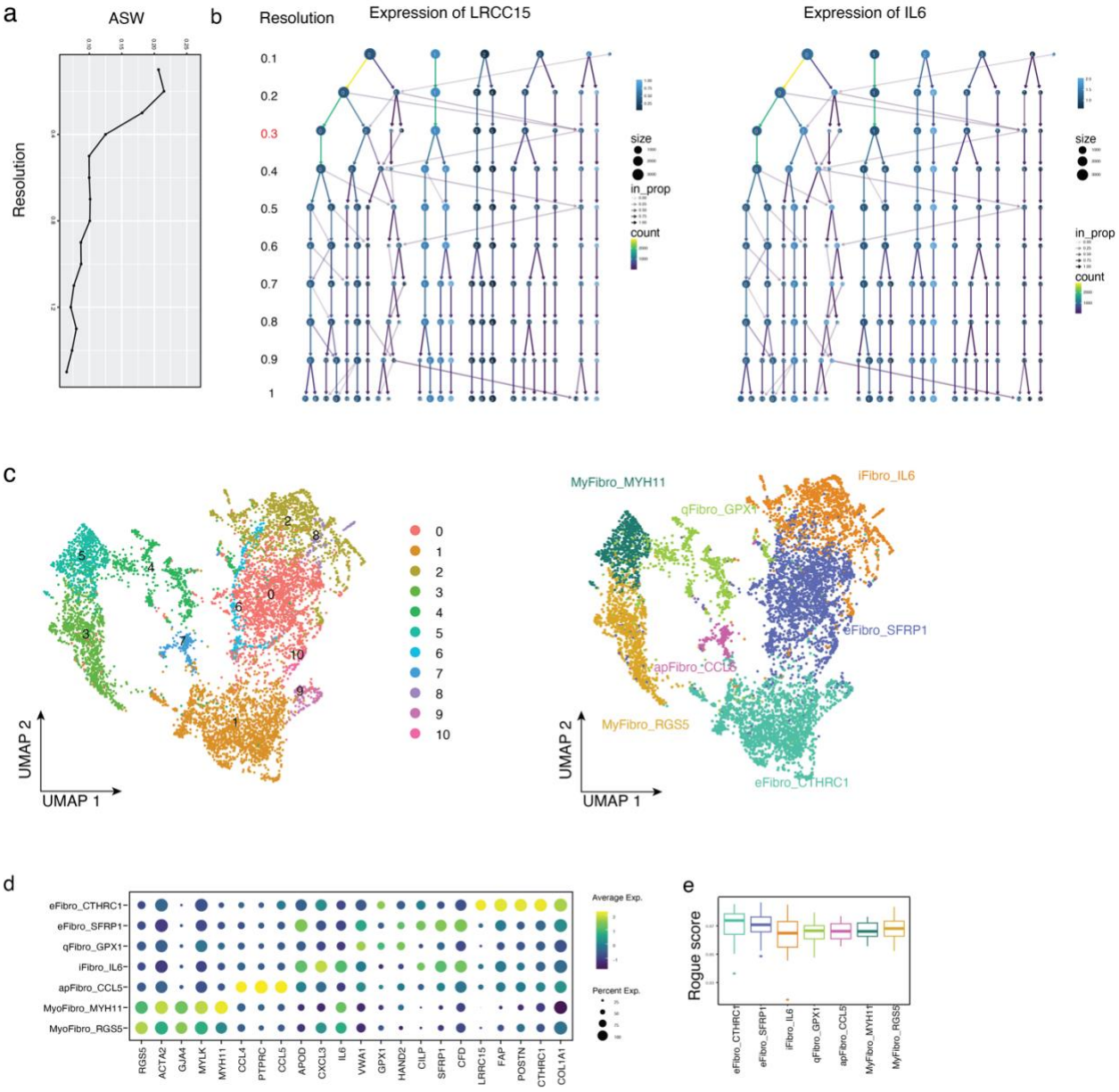
Then, based on the clustering results at the optimal resolution, we annotated the clusters based on the expression of marker genes. Subsequently, we employed ROGUE, an entropy-based universal metric, to evaluate cell type heterogeneity (Figure M1-6). ROGUE scores range from 0 to 1, providing an assessment of the purity of single-cell populations. One represents a completely pure subtype and zero represents the most heterogeneous state of a population. The annotated cell type with a median ROGUE value > 0.9, demonstrates that the cell type all are highly homogeneous. Conversely, a cell type with a median ROGUE score below 0.9 indicates significant heterogeneity, prompting further exploration to potentially re-annotate the cell types into more homogeneous subtypes with higher median ROGUE values.



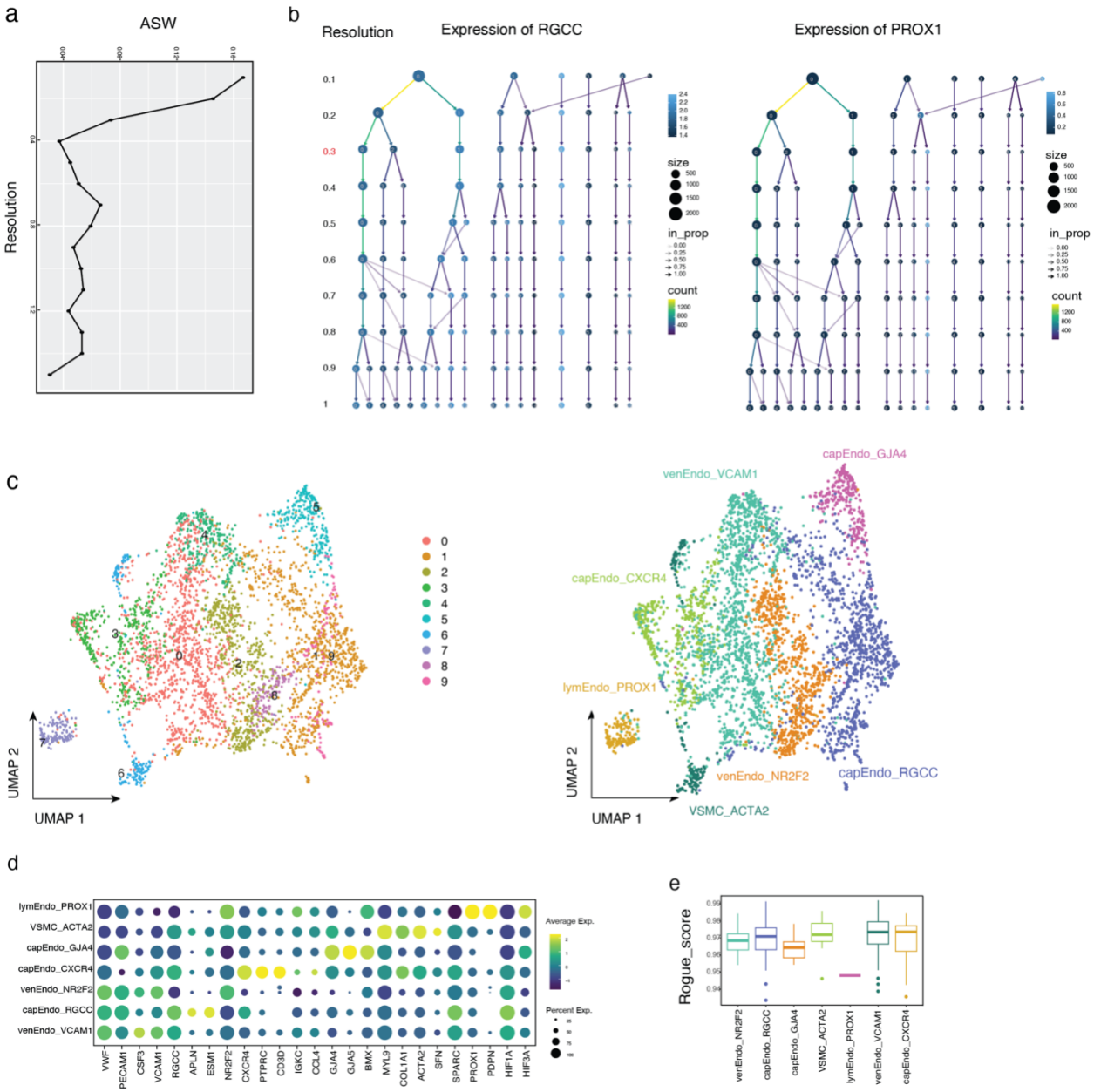
Supplementary Figure 1. Ascertain the optimal cluster resolutions and delineate cell type annotations for cytotoxic lymphocytes.
(a). Line plots depicting the average silhouette width of cytotoxic lymphocytes across a range of resolutions from 0.1 to 1.5. (b). Clustering trees of the cytotoxic lymphocytes colored according to the expression of known markers. The node colors indicate the average of the log2 TPM of samples in each cluster. *HAVCR2* signifies the exhausted state of cytotoxic lymphocytes, *FCER1G* highlights a subset of natural killer (NK) cells, and *CCR7* signifies the naïve state of cytotoxic lymphocytes. (c). UMAP visualization depicting the distribution of MetaCells of cytotoxic lymphocytes, with clusters displayed on the left and cell types on the right, each distinguished by a unique color. (d). Dot plot depicting the expression of representative marker genes of each cytotoxic lymphocytes subtypes. (e). Box plot illustrating cell purity for each cytotoxic lymphocyte cell type, calculated using ROGUE from 867 samples. The bottom of each box indicates the first quartile (Q1), and the top represents the third quartile (Q3). The height of the box reflects the interquartile range (IQR), and the horizontal line inside the box indicates the median. The whiskers extend to the positions of Q1 - 1.5 * IQR and Q3 + 1.5 * IQR.



Supplementary Figure 2. Ascertain the optimal cluster resolutions and delineate cell type annotations for myeloid cells. (a). Line plots depicting the average silhouette width of myeloid cells across a range of resolutions from 0.1 to 1.5. (b). Clustering trees of the myeloid cells colored according to the expression of known markers. The node colors indicate the average of the log2 TPM of samples in each cluster. *CD1C* identifies conventional type 2 dendritic cells (cDC2), *CLEC9A* shows a population of conventional type 1 dendritic cells (cDC1), and *C1QC* is a marker of macrophage cells. (c). UMAP visualization depicting the distribution of MetaCells of myeloid cells, with clusters displayed on the left and cell types on the right, each distinguished by a unique color. (d). Dot plot depicting the expression of representative marker genes of each myeloid subtypes. (e). Box plot illustrating cell purity for each myeloid cell type, calculated using ROGUE from 797 samples. The bottom of each box indicates the Q1, and the top represents the Q3. The height of the box reflects the IQR, and the horizontal line inside the box indicates the median. The whiskers extend to the positions of Q1 - 1.5 * IQR and Q3 + 1.5 * IQR.

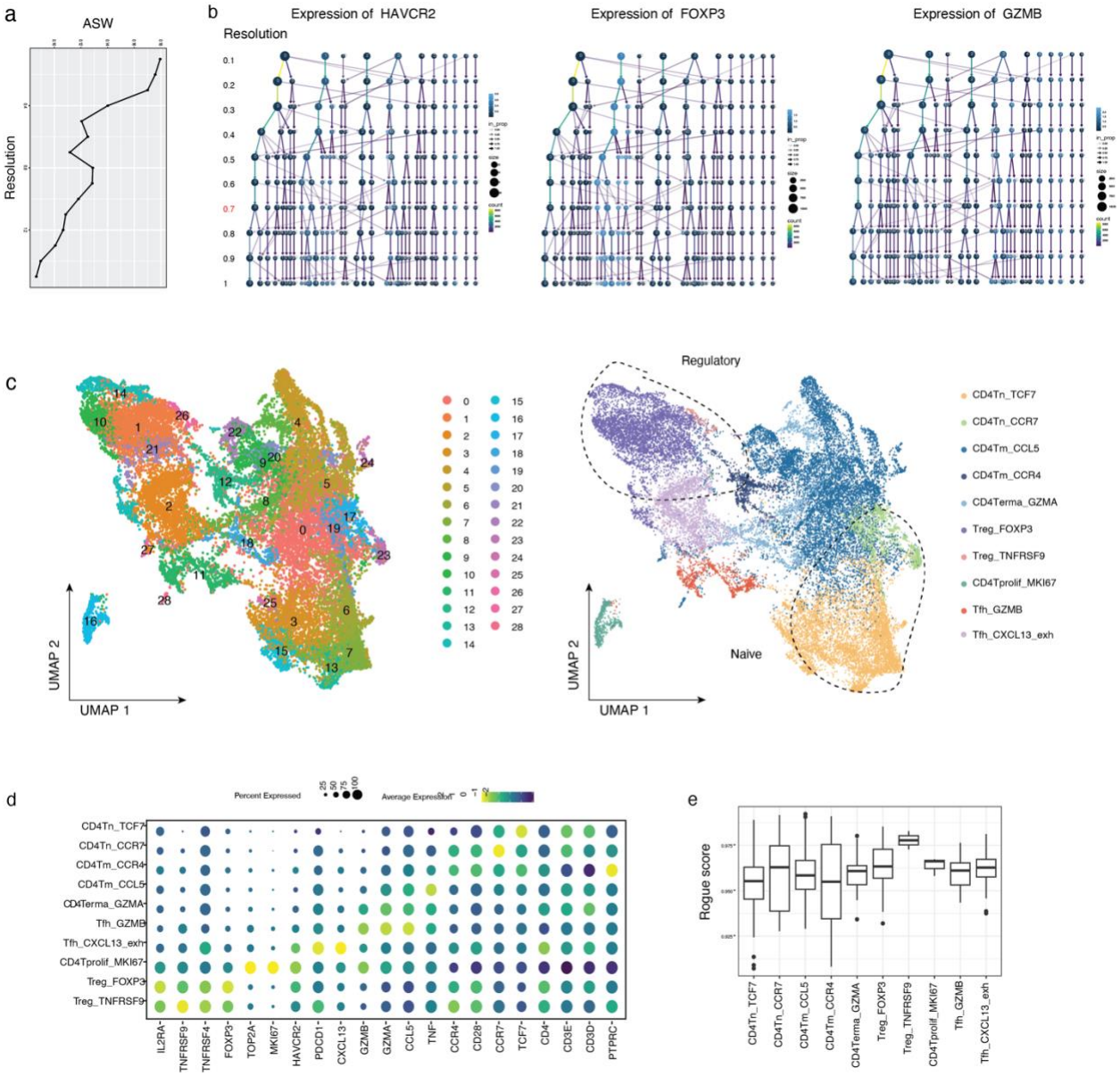


Supplementary Figure 3. Ascertain the optimal cluster resolutions and delineate cell type annotations for fibroblasts. (a). Line plots depicting the average silhouette width of fibroblasts across a range of resolutions from 0.1 to 1.5. (b). Clustering trees of the fibroblasts colored according to the expression of known markers. The node colors indicate the average of the log2 TPM of samples in each cluster. *LRRC15* distinguishes fibroblasts associated with extracellular matrix remodeling, while IL6 serves as a marker for immune-regulatory associated fibroblasts. (c). UMAP visualization depicting the distribution of MetaCells of fibroblasts, with clusters displayed on the left and cell types on the right, each distinguished by a unique color. (d). Dot plot depicting the expression of representative marker genes of each fibroblasts subtypes. (e). Box plot illustrating cell purity for each fibroblasts cell type, calculated using ROGUE from 379 samples. The bottom of each box indicates the Q1, and the top represents the Q3. The height of the box reflects the IQR, and the horizontal line inside the box indicates the median. The whiskers extend to the positions of Q1 - 1.5 * IQR and Q3 + 1.5 * IQR.

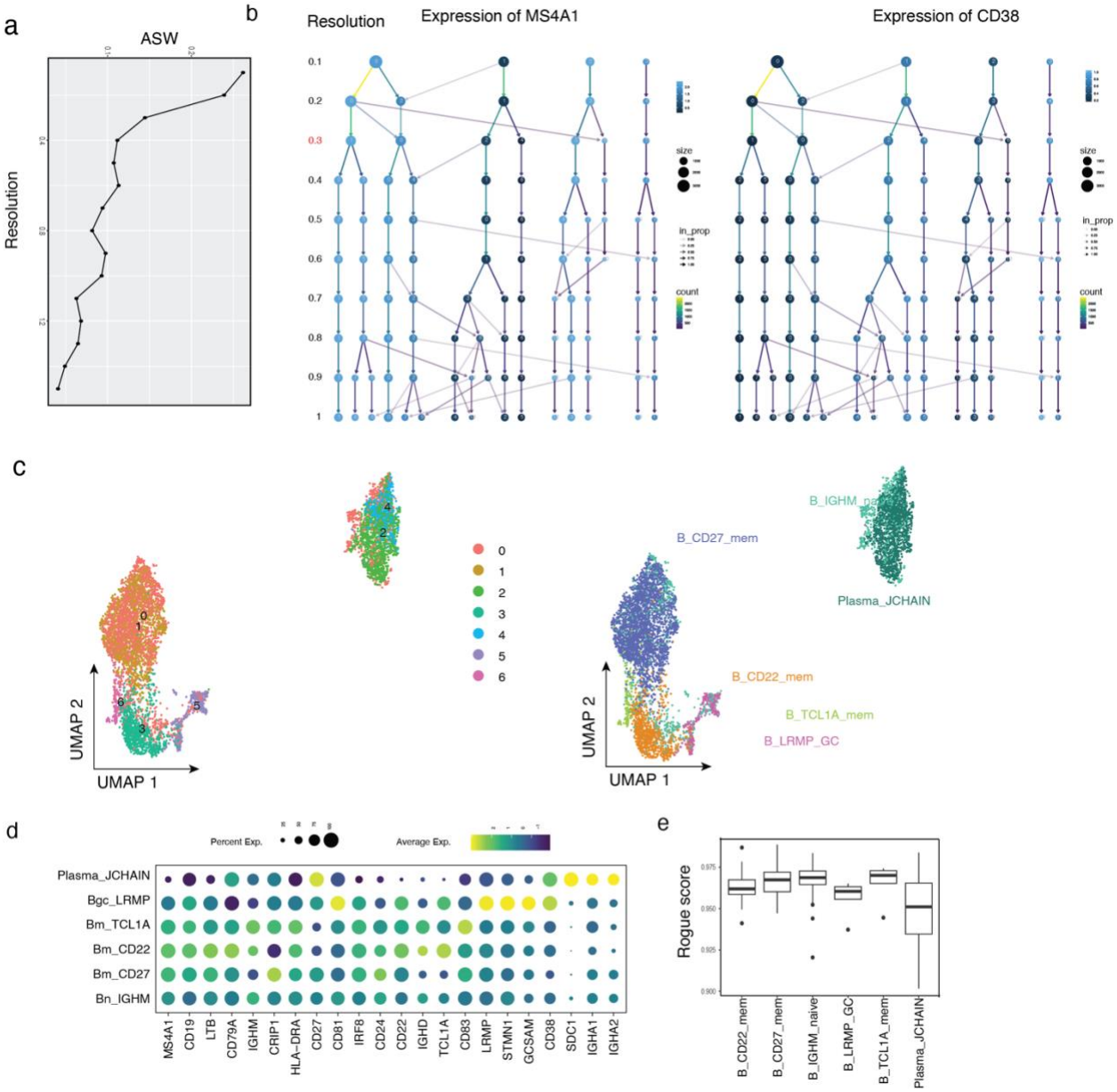


Supplementary Figure 4. Ascertain the optimal cluster resolutions and delineate cell type annotations for endothelial cells.

(a). Line plots depicting the average silhouette width of endothelial cells across a range of resolutions from 0.1 to 1.5. (b). Clustering trees of the endothelial cells colored according to the expression of known markers. The node colors indicate the average of the log2 TPM of samples in each cluster. *RGCC* identifies endothelial cells involved in tip generation, whereas *PROX1* functions as a marker for lymphatic endothelial cells. (c). UMAP visualization depicting the distribution of MetaCells of endothelial cells, with clusters displayed on the left and cell types on the right, each distinguished by a unique color. (d). Dot plot depicting the expression of representative marker genes of each endothelial subtypes. (e). Box plot illustrating cell purity for each endothelial cell type, calculated using ROGUE from 367 samples. The bottom of each box indicates the Q1, and the top represents the Q3. The height of the box reflects the IQR, and the horizontal line inside the box indicates the median. The whiskers extend to the positions of Q1 - 1.5 * IQR and Q3 + 1.5 * IQR.



Supplementary Figure 5. Ascertain the optimal cluster resolutions and delineate cell type annotations for conventional and regulatory lymphocytes.
(a). Line plots illustrating the average silhouette width of conventional and regulatory lymphocytes across a spectrum of resolutions ranging from 0.1 to 1.5. (b). Clustering trees of conventional and regulatory lymphocytes, colored based on the expression of established markers. Node colors indicate the average log2 TPM of samples within each cluster. *HAVCR2* denotes the exhausted state of conventional and regulatory lymphocytes, *FOXP3* highlights a subset of regulatory lymphocytes, and *GZMB* marks the cytotoxic state of conventional lymphocytes. (c). UMAP visualization showing the distribution of MetaCells for conventional and regulatory lymphocytes with clusters displayed on the left and cell types on the right, each distinguished by a unique color. (d). Dot plot demonstrating the expression of representative marker genes for each subtype of conventional and regulatory lymphocytes. (e). Boxplot displaying cell purity for each subtype of conventional and regulatory lymphocytes, calculated using ROGUE from 786 samples. The bottom of each box indicates the Q1, and the top represents the Q3. The height of the box reflects the IQR, and the horizontal line inside the box indicates the median. The whiskers extend to the positions of Q1 - 1.5 * IQR and Q3 + 1.5 * IQR.



Supplementary Figure 6. Ascertain the optimal cluster resolutions and delineate cell type annotations for B lymphocytes.
(a). Line plots illustrating the average silhouette width of B lymphocytes across a spectrum of resolutions ranging from 0.1 to 1.5. (b). Clustering trees of B lymphocytes, colored based on the expression of established markers. Node colors indicate the average log2 TPM of samples within each cluster. *MS4A1* highlights a subset of B cells, while *CD38* marks the activated state of B cells. (c). UMAP visualization showing the distribution of MetaCells for B lymphocytes, with clusters displayed on the left and cell types on the right, each distinguished by a unique color. (d). Dot plot demonstrating the expression of representative marker genes for each subtype of B lymphocytes. (e). Boxplot displaying cell purity for each subtype of B lymphocytes, calculated using ROGUE from 506 samples. The bottom of each box indicates the Q1, and the top represents the Q3. The height of the box reflects the IQR, and the horizontal line inside the box indicates the median. The whiskers extend to the positions of Q1 - 1.5 * IQR and Q3 + 1.5 * IQR.