To validate our observations, we examined publicly available scRNAseq datasets. We specifically selected studies that used 10X data from unbiasedly dissociated tumor samples to ensure proportional representation of each cell type.

In total, we retrieved 4 datasets from breast cancer (BC), 3 from colorectal cancer (CRC), 5 from hepatocellular carcinoma (HCC), and 5 from non-small cell lung cancer (NSCLC), encompassing 1,320,145 individual cells across 320 patient samples (BC, n=104; CRC, n=99; HCC, n=47; NSCLC, n=70). The full list of datasets used for this analysis can be found in Supplementary Table 4. For each cancer type, cell types were identified through marker gene annotation, while subclusters within each cell type were determined by label transfer based on our own pan-cancer atlas.

The heatmap below shows the normalized expression of marker genes for T-cell, B-cell, macrophage/monocyte, and DC subclusters across these publicly available datasets.