Sequencing data quality control

Raw sequencing data were processed to remove adapter sequences and low-quality regions using fastp (v1.0.1 Chen S, et al., 2018) to generate clean data for all subsequent analyses. The trimming procedure was performed as follows:

① Quality control was based on the nucleotide sequences located in the barcode and UMI regions of read1 and the probe region of read2.

② For tail-quality trimming, a stringent single-base sliding window was applied to the 3' end of each read. Any base with a quality score below 3 (using parameters --cut_tail_window_size 1 --cut_tail_mean_quality 3) triggered the truncation of that base and all subsequent bases in the read.

③ Adapter contamination was handled by enabling automatic adapter detection for paired-end reads

Processing the single cell RNA sequencing data

The SeekSoul Tools probe module (v2.0.0) was used for analyzing FFPE single-cell RNA-seq data. Clean reads were demultiplexed using the 17bp cell barcode and 12bp UMI embedded in Read 1, followed by adapter trimming and barcode validation. Read 2 was split into two 25bp segments: the first 25 bp (LHS) and the reverse complement of bp 26–50 (RHS), forming a pseudo-paired read set. These segments were aligned to the Chromium Human Transcriptome Probe Set (GRCh38) using Bowtie2 in local mode, requiring R1 to map in reverse and R2 in forward orientation, with a minimum alignment length of 23 bp. Only read pairs confidently mapping to the same probe were retained. UMIs were deduplicated using UMI-tools with the adjacency method (edit distance ≤ 1) for each barcode – probe combination. Probes marked as DEPRECATED_ were excluded; optionally, only probes with included=TRUE were retained. Cells were called using STAR solo with EmptyDrops_CR to distinguish real cells from background barcodes. The pipeline generated a filtered gene-cell count matrix and comprehensive QC metrics, including mapping rate, saturation, and median genes per cell. The final matrix was analyzed in Seurat v4 for log-normalization, PCA, clustering, and marker gene identification. Detailed QC reports were generated for each run.