Separate raw reads by specific cell barcode information attached in read 2 of the pair-ended reads.

Align UMI information to the corresponding read 1, then trim to remove the template switch oligo (TSO) sequence and polyA tail sequence.

Remove reads with adaptor contaminants and low-quality bases (N > 10%).

Align clean reads to the hg38 human transcriptome (UCSC) using TopHat (version 2.0.12).

Count uniquely mapped reads using htseq-count in the HTSeq package, and grouped them by cell-specific barcodes (Anders et al., 2014).

Remove duplicated transcripts with the same UMI sequences of each gene.

Count distinct UMIs of each gene (within a given individual cell) as the transcript copy number of that gene.

For all 2,579 sequenced single cells, we quantified the numbers of genes and transcripts in each cell. Cells with either fewer than 2,000 genes or fewer than 100,000 transcripts detected were filtered out. In total, 2,210 cells passed the filter standards. The expression levels were normalized by log2 (TPM/10+1), where TPM (transcripts-per-million) was calculated as the number of UMIs of each gene divided by all UMIs of a given cell, then multiplying by 1,000,000. We divided the TPM values by 10 because the UMI number of most of our single cell samples did not reach the order of 1,000,000 transcripts. Thus, we could avoid counting each transcript several times.