Feb 13, 2026

Whole exome sequencing (WES), mutation and copy number analysis and Single-cell RNA-seq (scRNA-seq)

DOI
https://dx.doi.org/10.17504/protocols.io.36wgq1y4xvk5/v1

Zhe Zhao[1], Guangli Suo[1], Xuan Xiong[1], Mengdie Yang[1]

[1]CAS Key Laboratory of Nano-Bio Interface, Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Sciences, Jiangsu, 215123, China.

MTS

Guangli Suo

DOI: https://dx.doi.org/10.17504/protocols.io.36wgq1y4xvk5/v1

Protocol Citation: Zhe Zhao, Guangli Suo, Xuan Xiong, Mengdie Yang 2026. Whole exome sequencing (WES), mutation and copy number analysis and Single-cell RNA-seq (scRNA-seq). **protocols.io** https://dx.doi.org/10.17504/protocols.io.36wgq1y4xvk5/v1

Protocol status: Working
We use this protocol and it's working

Created: February 13, 2026

Last Modified: February 13, 2026

Protocol Integer ID: 243166

Keywords: cell rna, rna, scrna, whole exome, mutation, seq

Funders Acknowledgements:
National Natural Science Foundation of China
Grant ID: 82202345

Abstract

Detailed protocols to the processes of whole exome sequencing and Single-cell RNA-seq.

Whole exome sequencing (WES), mutation and copy number analysis

1 The DNA library preparation and WES were outsourced to OEbiotech Corporation (Shanghai, China). Briefly, whole-genomic DNA libraries were prepared using the NEBNext Ultra II DNA library Prep Kit
from Illumina (New England Biolabs) following a protocol consisting of multiple enzymatic and purification steps. Human tissue DNA (GRCh37.p13) was utilized for filtering out background mutational noise, false positive mutations, and age-related mutations. After verification, the qualified DNA samples were randomly fragmented into 150 - 220 bp fragments by Covaris (M220). Agilent SureSelect
Human All Exon V6 was applied to build and capture library.
The Illumina NextSeq platforms with sequencing depth greater than 100× were utilized for WES to investigate genomic mutations and copy number alterations. A proprietary bioinformatics pipeline was utilized to process the genomic data, enabling identification of various classes of genomic abnormalities. Specifically, we employed the BWA algorithm from the HaplotypeCaller module of GATK4 software to compare the reference genome with samples, facilitating detection of small insertions/deletions (InDel) and single-nucleotide substitutions (SNPs). To enhance mutation detection accuracy, base quality recalibration was performed using known InDel and SNP databases along with the BaseRecalibrator module of GATK4 software. Only mutation sites with QD values exceeding 2, which is calculated by dividing the variation quality value by coverage depth, were retained to minimize error rates in InDel and SNP detection. Subsequently, InDel and SNP results were annotated against Refseq and other databases including EXAC, esp6500, gnomAD, SIFT, clinvar, PolyPhen, MutationTaster, COSMIC, gwasCatalog and OMIM using Annovar software. Copy number variation (CNV), representing changes in copy numbers of genome fragments, constitutes an essential component of genomic Structural variation (SV), mainly encompassing deletion and duplication of genomic segment. CNV can lead to the development of complex diseases such as cancer; therefore chromosome-level deletion and amplification analysis has emerged as a focal point in tumor research. We employed Control-FREEC software to detect somatic CNVs in both primary tissues and tumor paired MTSs.
Generally, mutations in cancer cells can be classified into two types: driver mutations and passenger mutation. The genomic landscape of clinical BRC tissue or MTSs was summarized and visualized using R package ‘maftools’. The mutation patterns were fitted using R package ‘Mutational Patterns’ to identify components matching to the 96 COSMIC mutational signatures (V.3.2). To avoid signature misattribution, a refitting procedure was performed with 1,000 bootstrap iterations. The relative contribution of each COSMIC signature for each sample was then plotted using the R package ‘ggplot2’.

Single-cell RNA-seq (scRNA-seq)

2 Patient-derived MTSs and their corresponding original BRC tissues from patient (P136) were used for scRNA-seq analysis in this study. The single-cell transcriptome libraries were constructed using the 10x Genomics platform. The scRNA-seq libraries were prepared with the Chromium Next GEM Single Cell 3' Library and Gel Beads Kit v3.1 (PN-1000121) according to the manufacturer's instructions. Quantification of cDNA and DNA libraries for each sample was performed using Qubit 4 Fluorometers (Thermo Fisher Scientific), and quality control was conducted with the Agilent 2100 Bioanalyzer (Agilent Technologies). All single-cell transcriptomes were sequenced on the Illumina Novaseq 6000 platform conducted by OE Biotech Co., Ltd. (Shanghai, China), generating paired-end reads with a length of 150 bp.

After sequencing, the scRNA-seq data underwent processing and clustering. The CellRanger version 6.0.1 (10x Genomics) was utilized for aligning reads to the human reference genome (GRCh38), and the raw count matrix for each sample was obtained from the CellRanger unique molecular identifier (UMI) matrix output. Genes expressed in at least 0.1% cells were retained for downstream analyses. Initially, the percentage of counts originating from mitochondrial RNA and heat shock-related RNA per cell was calculated first to ensure high-quality datasets. Subsequently, cells were filtered based on higher-quality characteristics, including mitochondrial reads below 10%, a number of detected genes ranging from 400 to 8,000, and a number of UMIs ranging from 500 to 50,000. Python package Scrublet90 was applied to estimate potential doublets in each sample with the expected doublet rate of 5%. Standard preprocessing steps involved normalizing feature expression measurements for each cell by total expression, followed by multiplication with a scale factor (1e4), and log transformation of the result. Subsequently, we scaled expression values regressing out the percent of mitochondrial counts, count numbers, detected genes and heat shock-related genes. We selected the top 5,000 most variable genes for downstream analyses. Principal component analysis (PCA) was performed as a linear dimensionality reduction on the scaled data with 50 principal components retained. To correct the batch effects arising from different samples, we applied BBKNN91 to generate a batch-balanced k nearest neighbor (KNN) graph with parameters neighbors_within_batch = 3.

Subsequently, we employed the Leiden algorithm to iteratively cluster cells together, initially setting the resolution at 0.2. We identified the major 8 cell clusters based on well-established cell markers. For visualization purposes, Manifold Approximation and Projection (UMAP) was applied to the KNN graph using a Python-based kernel. Furthermore, subpopulations within each major cell cluster were identified following the same aforementioned procedure starting from the unfiltered UMI matrix. To obtain different fine-grained clustering results, we varied the parameter resolution in the Louvain algorithm from 0.3 to 0.9 for each sub-clustering analysis. All clustering analyses were performed using Scanpy (version 1.8.1), a Python-based toolkit.

To annotate subpopulations within each major cell cluster, we initially generated the normalized SCT matrix from UMI count data using SCTransform, an R-based toolkit provided by Seurat (version 4.0.5). Subsequently, differentially expressed genes (DEGs) were identified on the SCT matrix using FindAllMarkers (MAST test with Bonferroni correction for multiple testing; adjusted $P < 0.05$). We exclusively considered genes that were detected in at least 10% of the cells within the cluster and exhibited an average fold difference of at least 0.25-fold (log-scale) between the cells in the cluster and all other cells. The cell cluster identities were determined by top-ranked DEGs and previously reported biologically related genes. All single-cell RNA-seq data that support the finding of this study has been deposited at SequenceRead Archive (SRA) under accession number PRJNA1165172.