FEB 07, 2024

DOI:
dx.doi.org/10.17504/protocols.io.j8nlkoq25v5r/v1

Protocol Citation: Peter Kilfeather 2024. Processing of raw Stereo-seq data, quality control and cell type identification. **protocols.io** https://dx.doi.org/10.17504/protocols.io.j8nlkoq25v5r/v1

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

Created: Feb 07, 2024

Processing of raw Stereo-seq data, quality control and cell type identification

In 1 collection

Peter Kilfeather[1]

[1]University of Oxford

PK

Peter Kilfeather

ABSTRACT

Processing of raw Stereo-seq data, quality control and cell type identification methods from Kilfeather, Khoo et al., 2024

Last Modified: Feb 07, 2024

PROTOCOL integer ID: 94801

Funders Acknowledgement:
Aligning Science Across Parkinson's
Grant ID: ASAP-020370
Monument Trust Discovery Award from Parkinson's UK
Grant ID: J-1403

Protocol

1 Raw gene-by-spot data per sample were aggregated to create a 2-dimensional image of RNA signal for each sample using custom Python (v3.9.0, RRID:SCR_008394) scripts. To segment individual cells, each image was subjected to a processing pipeline written in Python (workflow illustrated in Supplementary Figure 1). Steps taken: Mask generation from unspliced counts: Gaussian filter (sigma = 5), background subtraction (white tophat, 50 pixels), Otsu thresholding, conversion to mask, fill holes, watershed. Mask generation from spliced + unspliced counts: As for unspliced counts. Watershed boundaries from the spliced + unspliced mask were then subtracted from the unspliced mask. Objects were retained from the spliced + unspliced mask that overlapped with the unspliced mask. Each cell was labelled using the label function in SciPy (v1.9.0, RRID:SCR_008058). Raw gene-by-spot data were then aggregated to the gene-by-cell level and imported into scanpy (v1.9.1, RRID:SCR_018139)62.

The initial Stereo-seq dataset contained 497,766 cells. In a first round of filtering, low complexity and putative doublet/triplet cells were filtered. To remove low complexity cells, a minimum gene detection cutoff of 200 was selected. To remove putative doublet/triplet cells, a maximum gene detection cutoff was set on a per-brain basis to the median number of genes detected + 4 median absolute deviations. After first round filtering, 415,402 cells remained. Count data were subsequently processed using SCVI, including sample brain of origin as a categorical covariate, and the number of genes detected per cell as a continuous covariate78. A uniform manifold projection (UMAP) was generated for cells from each mouse brain and overlaid, showing a similar pattern of separation (Supplementary Figure 1A).

Cell type identification was performed in two rounds. In a first pass, leiden clustering was performed at iteratively greater resolutions79. Cell types without clear spatial organization e.g. some glial types, were annotated based on marker gene enrichment. Remaining cells were then processed using SEDR in order to include spatial information in the clustering process80. Mclust was used for clustering spatially defined cell types81. If a cluster contained fewer than 200 cells, it was considered final and no further subclustering was performed. Dopaminergic neurons were labelled according to their anatomical region (SN and VTA), based on their spatial coordinates.

For differential expression analysis in Stereo-seq samples, gene counts were normalized to the total number of counts per cell and log-transformed. Marker genes for each cluster were identified using the Wilcoxon rank-sum test, providing two-sided P values. The identity of each cell type was annotated by integrating marker gene data with previous literature and by confirming the spatial distribution of clustered cells.

A second round of cell filtering was performed after clustering to remove cells that were enriched in both Snap25 and Plp1. We suspected that these cells represent neuronal/glial contaminated mixtures, as previously reported in single cell data82. After second-round filtering, 355,307 cells remained.