[BIOINFO 2025]

Somatic Variant Calling in snRNA-seq Metacells Reveals Clonal Phylogeny

Junseong Bae[1], Jeongbin Park[1,2*]

[1] *Interdisciplinary Program of Genomic Data Science, Pusan National University*
[2] *School of Biomedical Convergence Engineering, Pusan National University*
**Corresponding author: jeongbin.park@pusan.ac.kr*

Single-cell and single-nucleus RNA-seq provide rich phenotypes but are typically shallow for variant discovery. We present a streamlined pipeline that boosts signal for somatic SNV/indel calling by aggregating local neighborhoods into "metacells", then projects variants back to individual cells to reconstruct clonal structure. Starting from splicing-aware alignments, we (i) build a kNN graph and form ~100–500 metacells per dataset via randomized neighborhood pooling; (ii) perform joint multi-sample variant calling per metacell; (iii) map per-variant counts back to cells and create a binary variant×cell matrix; and (iv) infer subclonal phylogeny and trajectories aligned with transcriptional manifolds. Across public tumor snRNA-seq datasets, metacell pooling improves sensitivity for known driver mutations and detects rare subclones that single-cell callers miss at native read depth. False positives are reduced using strand-bias checks and RNA-editing masks. Metacell pooling increases effective depth ~10–100x without altering cell-state boundaries, enabling robust localization of tumor-restricted mutations on UMAPs that delineate malignant clusters. Based on the variant calling in metacells, we were able to infer subclonal phylogeny and lineage trajectories across malignant transitions. All notebooks and data required to reproduce the above results are available on our GitHub page.