Bento Tools: Subcellular Analysis of Spatial Transcriptomics Data

Aastha Punjabi (210017)

1 Problem Illustration

Understanding subcellular RNA organization is essential for deciphering cellular functions and disease mechanisms. Conventional spatial transcriptomics methods lack resolution at the subcellular level, focus- ing instead on multicellular or tissue scales. Bento bridges this gap by providing a suite of Python tools for subcellular RNA analysis, enabling researchers to explore RNA localization, gene-gene colocalization, and spatial domain identification at single-molecule resolution.

2 Implementation Details

- **RNAforest**
  - **Goal:** Annotate RNA localization patterns in subcellular compartments
  - **Methodology:** RNAforest uses a multilabel random forest classifier trained on features derived from cellular and nuclear boundaries. These features describe spatial distributions, including proximity to cellular landmarks and density.
  - **Output:** Classification of RNA localization patterns for each gene in every cell.
- **RNAcoloc**
  - **Goal:** Calculate context-specific gene colocalization within cellular compartments.
  - **Methodology:** RNAcoloc uses the Colocation Quotient (CLQ) to measure gene colocalization in distinct cell compart- ments (e.g., nucleus, cytoplasm). Tensor decomposition (PARAFAC) identifies patterns of colocalization across cells and compartments.
  - **Output:** Compartment-specific gene colocalization scores, revealing spatial interaction networks.
- **RNAflux**
  - **Goal:** Identify and quantify transcriptionally distinct subcellular domains.
  - **Methodology:** RNAflux calculates local RNA composition vectors and uses self-organizing maps (SOMs) to cluster pixels into subcellular domains.
  - **Output:** Visual representation of subcellular regions, showing spatial variation in RNA localization.

3 Dataset Description

The toolkit was evaluated using multiple spatial transcriptomics datasets:

- **MERFISH dataset (U2-OS cells):** Captures 130 genes in human osteosarcoma cells with high molecule density (111 molecules per gene per cell on average).
- **seqFISH+ dataset (3T3 cells):** Contains a broader gene panel (10,000 genes) but lower detec- tion efficiency (8 molecules per gene per cell).
- **Cardiomyocyte Dataset (iPSC-derived):** Generated with Molecular Cartography, measuring 100 genes crucial for cardiomyocyte function. Used to analyze doxorubicin-induced RNA localiza- tion shifts.

4 Results

4.1 RNA Localization Patterns (RNAforest)

Findings: RNAforest classified genes into distinct localization patterns, successfully identifying nuclear, cytoplasmic, and cell-edge localization types.

4.2 Compartment-Specific Colocalization (RNAcoloc)

Findings: RNAcoloc highlighted compartment-specific colocalization patterns, with nuclear RNAs showing tighter colocalization compared to cytoplasmic RNAs.

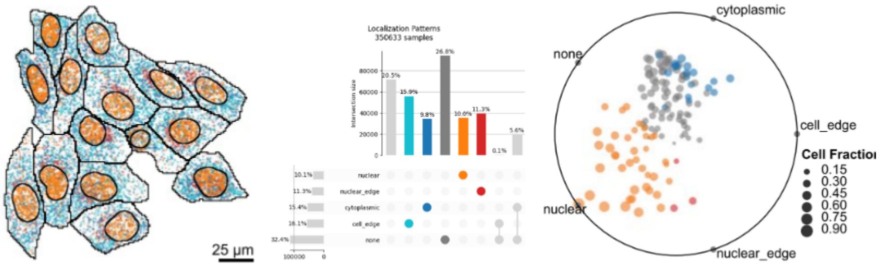

Figure 1: RNA Localization Patterns in U2-OS cell. Visualizes nuclear, cytoplasmic, and cell-edge RNA patterns.

Class	Best Threshold	Best F1-Score	Default Threshold	Default F1-Score
1	0.453	0.9528535980148883	0.5	0.9515527950310559
2	0.434	0.7661691542288558	0.5	0.7503267973856209
3	0.379	0.7907514450867051	0.5	0.7636363636363637
4	0.437	0.9701492537313433	0.5	0.9660377358490566
5	0.505	0.9749373433583959	0.5	0.9737171464330413

Table 1: Performance comparison of Best and Default Thresholds for each Class

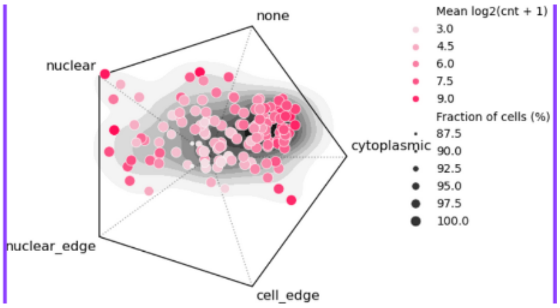

Figure 2: Gene distribution in Subcellular Regions

Gene Pair	Nuclear CLQ	Cytoplasmic CLQ
PIK3CA - DYNC1H1	0.85	1.24
MALAT1 - CNR2	1.09	0.76
SOD2 - FBN2	0.92	0.83

Table 2: Compartment-Specific Gene Colocalization in U2-OS Cells

4.3 Subcellular Domains (RNAflux)

Findings: RNAflux detected transcriptionally distinct subcellular regions in cardiomyocytes. Under doxorubicin treatment, RNA was significantly depleted from the endoplasmic reticulum, indicating stress-induced changes.

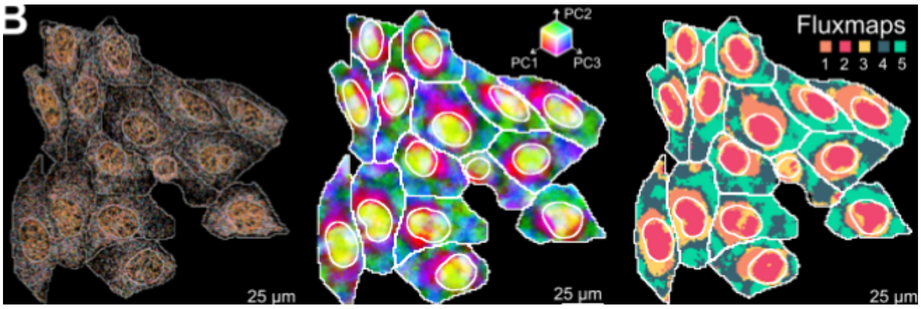

Figure 3: RNAflux Analysis of Subcellular Domains in U2-OS Cells.

5 Adaptive Sampling Algorithm for RNAforest

5.1 Algorithm Changes and Enhancements

Introduction of a Two-Phase Sampling Process:

- **Phase 1**: Initial Screening to classify regions based on complexity (RNA density and spatial variability).
- **Phase 2**: Progressive Sampling with differential sampling intensity based on the complexity classification from Phase 1.

Classification of Complexity:

- **Low-Complexity Regions** are identified as areas with low RNA density or uniform spatial patterns.
- **High-Complexity Regions** are identified as areas with high RNA density or irregular spatial patterns, often near cellular landmarks (e.g., cell edges, nuclear boundaries).

Sampling Strategy Adjustments:

- **Dense Sampling in High-Complexity Regions**:
  - Increase sampling density in high-complexity regions, enabling more detailed feature extraction.
  - Calculate spatial features (proximity, symmetry, dispersion) more intensively in these regions.
- **Sparse Sampling in Low-Complexity Regions**:
  - Use sparse sampling to minimize computational load in low-complexity regions.
  - Only essential spatial features are calculated, as intricate patterns are unlikely here.

5.2 Implementation in RNAforest

Grid-Based Region Division:

- Each cell is divided into a grid, and initial RNA density metrics are calculated for each grid region.
- This preliminary step requires minimal computation and helps identify areas where detailed sampling is most valuable.

Dynamic Sampling and Feature Calculation:

- A conditional sampling function adapts sampling density based on the complexity of each region.
- High and low-complexity areas regions apply a detailed and a reduced feature set respectively.

Integration with the Random Forest Classifier:

- Feature vectors generated from dense and sparse sampling are combined and passed to RNAforest classifier.
- Adaptive sampling output is treated uniformly by the classifier, ensuring no changes are needed in the model's structure.

5.3 Summary

This adaptive approach **optimizes computational resources** by concentrating analysis on high-density RNA regions where patterns are more complex and significant, reducing processing time without sacrificing accuracy. By implementing grid-based region classification and conditional sampling, RNAforest gains both efficiency and scalability, allowing it to handle larger datasets and capture more detailed RNA localization patterns in complex cell areas.