Organoid Fidelity via Multi-Omics Integration

Project: Organoid-Data-Analysis
Notebook: nb4_multiomics
Dataset: GSE75140 (Camp et al. 2015, *Cell*) + synthetic protein panel

---

Executive Summary

This notebook demonstrates how integrating RNA with a protein modality improves the assessment of organoid fidelity. Using cerebral organoid data (GSE75140) and a synthetic 30-marker panel, we show that joint factors highlight progenitor, neuronal, and glial programs more clearly than RNA alone. A shuffle baseline confirms that protein contributes genuine structure, while scorecards reveal heterogeneity across organoid clusters. Together, these results provide a reproducible framework and guardrails for multi-omic fidelity analysis.

Purpose

Evaluating organoid fidelity means asking a difficult question: *do organoids truly reproduce human developmental programs, or only parts of them?*

RNA profiles alone give us one answer — they reveal which transcripts are present. But RNA does not always match protein abundance or activity. Post-transcriptional regulation, alternative splicing, and translational control all affect how faithfully RNA predicts biology.

This notebook shows how adding a second modality, even a synthetic one, changes the interpretation of fidelity. By integrating RNA with a curated protein panel, we uncover latent factors that separate progenitors, neurons, and glial programs more clearly than RNA alone. The workflow also embeds guardrails — shuffle baselines, block scaling, and low-N exclusions — to make results reliable and reproducible.

Challenges in Organoid Fidelity

Fidelity is **multi-dimensional**: transcriptional, proteomic, and epigenetic. Defining it requires explicit choices. Should organoids be judged by how closely they resemble fetal tissue? By whether they diverge from adult tissue? By their reproducibility across replicates? Each answer frames fidelity differently.

Integration across modalities is **technically challenging**. RNA captures thousands of features, while proteins or ATAC capture tens to hundreds. Without scaling, RNA dominates the analysis. Our block-scaling approach equalizes contributions, preventing misleading

results.

Datasets remain **RNA-heavy**. Multi-omics data for organoids are scarce, so most fidelity assessments use transcriptomes only. This notebook uses a synthetic protein panel to illustrate the framework, but future work must rely on true multi-modal assays (e.g. CITE-seq, RNA+ATAC).

Organoids are **internally heterogeneous**. A single organoid may contain progenitors, immature neurons, and glial precursors. Fidelity measured at the organoid level hides this diversity. Fidelity must be quantified at the cluster or subpopulation level, as shown here with organoid × cluster scorecards.

Latent factors are **abstract without interpretation**. Factor 1 or Factor 2 is just math until linked to SOX2, MAP2, or GFAP. Mapping loadings back to genes and proteins anchors factors in biology.

Framework for Analysis

The notebook follows a modular workflow:

1. **Setup** — environment, paths, helper functions.
2. **Load data** — GSE75140 RNA and a synthetic 30-marker protein panel.
3. **Preprocess modalities** — RNA HVGs → PCA; Protein scaling → PCA.
4. **Integrate** — block-scaled concatenation → Factor Analysis.
5. **Cluster** — Leiden on latent factors to define subpopulations.
6. **Embed** — UMAP on factors for visualization.
7. **Variance explained** — per factor and per modality.
8. **Interpret factors** — map loadings to genes/proteins; plot heatmaps.
9. **Robustness baseline** — shuffle Protein to confirm contribution.
10. **Scorecards** — summarize factor profiles per organoid and per cluster, with low-N exclusions.

Each step is documented with code, outputs, and short interpretation blocks so the analysis can be followed as a tutorial.

Guardrails

- **Fixed seed (1337):** reproducibility for PCA, FA, and UMAP.
- **Block scaling:** ensures RNA and Protein contribute fairly.
- **Shuffle baseline:** validates that Protein adds true signal.
- **Low-N cutoff (25 cells):** prevents over-interpreting small groups.

Key Takeaways

RNA alone cannot fully capture organoid fidelity. With protein integration, factors align with recognizable neurodevelopmental programs.

- **Progenitors:** SOX2, PAX6, NES.
- **Neurons:** MAP2, DCX, NEUROD1.
- **Glia:** GFAP, OLIG2, S100B.

Protein makes a measurable contribution. In the integrated model, it explains ~21% of variance; after shuffling, this drops to ~6%. This proves the second modality adds genuine structure, not noise.

Fidelity is heterogeneous. Scorecards show organoid × cluster units differ in factor profiles: some progenitor-like, others neuron- or glia-like. Fidelity cannot be measured as a single score for the entire organoid.

Together, these results show that multi-omic integration is not just a technical exercise — it fundamentally changes how we interpret organoid fidelity.

Next Steps Beyond This Notebook

- **Pathway enrichment:** link factor loadings to Reactome or GO terms.
- **Reference comparisons:** benchmark organoid factors against fetal and adult tissue.
- **True multi-omics:** apply the workflow to real RNA+Protein or RNA+ATAC datasets.
- **Cross-condition studies:** test fidelity across labs, protocols, or genetic backgrounds.