Organoid Heterogeneity

Project: Organoid-Data-Analysis
Notebook: nb2_heterogeneity
Dataset: GSE75140 (Camp et al. 2015, *Cell*)

Purpose

Organoids are powerful models of human development and disease, but they are also **heterogeneous**.

- *Within organoids*: multiple neural lineages coexist (radial glia, intermediate progenitors, neurons).
- *Between organoids*: replicates vary in how much of each lineage they generate.
- *Across conditions*: protocols, batches, or genetic backgrounds add further variability.

This notebook demonstrates how to **quantify and visualize heterogeneity** in single-cell data from cerebral organoids, using GSE75140 as a case study. All instructions, annotations, interpretations, and saved artifacts are documented directly in the notebook.

Challenges in Organoid Heterogeneity

1. **Replicate variability** — even under the same protocol, organoids differ in lineage proportions.
   Example: Camp et al. showed basal progenitors (TBR2^{+}/EOMES^{+}) are inconsistently represented.
2. **Sampling depth** — most organoids contribute only a handful of cells, inflating noise.
   Example: An organoid with 3 neurons looks homogeneous but is simply under-sampled.
3. **Interpretation pitfalls** — embeddings rarely separate organoids; heterogeneity is often compositional.
   Example: Radial glia from different organoids intermix in UMAP, masking variability visible only in proportions.
4. **Reproducibility** — need to separate true biological variability (lineage shifts, maturation) from technical effects.

Framework for Analysis

This notebook structures heterogeneity analysis into five layers:

1. **Composition** — how much of each state an organoid contains.
   Example: Some organoids skew neuron-heavy, while others retain more progenitors.

2. **Diversity** — how balanced internal mixtures of states are.
   Example: A high-diversity organoid may resemble fetal tissue mixtures; low-diversity ones collapse onto a single lineage.

3. **Between-organoid dissimilarity** — how far apart organoids are in lineage balance.
   Example: Neuron-rich vs progenitor-rich organoids form distinct branches in distance maps.

4. **Embedding concordance** — whether organoids overlap in UMAP/PCA space.
   Example: Neurons from different organoids intermix, suggesting conserved transcriptional programs.

5. **Intra-type heterogeneity** — whether organoids differ *within* a lineage.
   Example: Neurons mix well, but progenitors sometimes diverge in cycling state or maturation.

Guardrails:

- **Random-mixing baselines** tell us what "good mixing" means given organoid sizes.
- **Low-N thresholds** prevent over-interpreting small organoids.
- **Pseudobulk profiles** stabilize variance for robust comparisons.

Steps in This Notebook

1. **Setup** — initialize environment, seed, paths, helper functions.
2. **Load Data** — import counts, orient cells × genes, assign organoid_id.
3. **QC (document only)** — visualize counts/genes/mito metrics; no filtering applied.
4. **Normalize → HVGs → PCA** — standardize counts, select variable genes, run PCA.
5. **Neighbors, UMAP, Leiden clustering** — embed cells and call clusters as proxy states.
6. **Composition by organoid** — quantify cluster proportions per organoid.
7. **Diversity indices** — compute Shannon/Simpson/evenness per organoid.
8. **Between-organoid dissimilarity** — compare organoid compositions via Jensen–Shannon distance.

9. **Embedding concordance** — measure organoid mixing in UMAP (neighbors, entropy, silhouette).
10. **Intra-type heterogeneity** — repeat mixing analysis within each lineage.
11. **Baselines & guards** — compute random-mixing expectations; flag low-N organoids.
12. **Interpretation summary** — bullet-point synthesis of results.
13. **Save artifacts + session info** — export processed data, metrics, and run environment.

Addressing Organoid Heterogeneity — Broader Perspective

Multiple Levels of Heterogeneity

- **Within organoid**: coexistence of radial glia, progenitors, and neurons in a single replicate.
- **Between organoids**: replicate-to-replicate variation in lineage balance (e.g., one organoid is neuron-rich, another progenitor-rich).
- **Across conditions**: protocol- or line-specific differences layered on top.

What to Prioritize

- **Composition and diversity** are the most robust indicators.
- **Embedding concordance** is mainly a sanity check; organoids often overlap in shared state space.
- **Within-type analysis** can uncover subtle differences in progenitor cycling or neuronal maturation.

Guardrails

- Compare observed mixing to **expected random baselines** to contextualize values.
- Apply **low-N filters** so outliers are not over-interpreted.
- Use **pseudobulk (organoid × type)** for stable comparisons and DE analysis.

Biological Context

Heterogeneity is not mere noise — it is a **defining feature** of organoids.

- Some variability reflects human developmental stochasticity (different progenitor-to-neuron ratios).

- Some arises from technical or sampling effects.
  Example: Quadrato et al. (2017, *Nature*) showed organoids reproducibly generate neurons, but neuronal maturation and activity signatures vary substantially across replicates.

How This Notebook Helps

- Provides a stepwise framework to separate composition-driven from state-driven heterogeneity.
- Embeds baselines and guardrails for interpretation.
- Offers a tutorial template applicable to other datasets (e.g., fetal cortex, disease models).

Next Steps Beyond This Notebook

- Map organoid heterogeneity against fetal cortical references.
- Apply the framework in perturbation studies (does a treatment increase or reduce heterogeneity?).
- Extend to multi-modal data (transcriptomic, epigenomic, spatial) to uncover whether heterogeneity lies in state identity, lineage potential, or tissue architecture.

Key Takeaways

- **Core states are reproducible**: organoids consistently generate radial glia and neurons.
- **Heterogeneity is compositional**: replicate variability lies mainly in proportions of progenitors vs neurons.
- **Shared manifolds**: organoids overlap in embeddings; lineage programs are conserved.
- **Guardrails are essential**: small organoids exaggerate noise; baselines provide context for interpretation.
