Organoid scRNA-seq: QC & Batch Correction (Human Intestinal Organoids)

Dataset: GSE156760 — human intestinal organoids (colon & ileum), conditions: mock, 12 hpi, 24 hpi (10x scRNA-seq).

Introduction

This notebook focuses on **quality control and batch correction** for single-cell RNA-seq from human intestinal organoids. The aim is a clean, comparable dataset across samples and an integrated representation that removes technical sample effects while **preserving biological structure** (organ and timepoint). Light, contextual checks (cell-type markers, interferon response, optional reference comparison) verify that biology remains intact after correction.

All instructions, annotations, interpretations, and saved artifacts are documented directly in the notebook.

What this demonstrates (short story)

- **Quality assessment:** Per-sample thresholds (mitochondrial/ribosomal content, genes, UMIs), cell-cycle scores, doublet filtering, and an ambient-RNA heuristic produce clean inputs.
- **Integration that preserves biology:** Harmony reduces sample effects while organ and timepoint structure remain; this is summarized by **cross-batch fraction (CBF)** and **silhouette** metrics with pre/post UMAPs for context.
- **Biology check:** Marker DotPlot and simple signatures outline expected epithelial lineages; an interferon-stimulated gene (ISG) signal increases at 12/24 hpi; (optional) CellTypist provides an external reference comparison.

Step-by-step outline (consistent with the notebook)

1. **Ingest & standardize** — load GEO matrices; make gene/barcode names unique; add `sample_id`, `organ`, `timepoint`.
2. **QC metrics** — compute `pct_counts_mt`, `pct_counts_ribo`, `n_genes_by_counts`, `total_counts`; score cell cycle.

3. **Per-sample filtering** — apply IQR-based thresholds with caps/floors; retain high-quality cells.
4. **Doublet handling** — predict with Scrublet per sample; remove flagged doublets.
5. **Ambient heuristic** — report a simple contamination indicator per sample.
6. **Normalize & log** — library-size normalization; `log1p`; preserve `.raw`.
7. **HVGs & PCA (pre)** — select batch-aware HVGs; compute PCA; neighbors/UMAP (baseline, pre-integration).
8. **Batch correction (Harmony)** — integrate on PCA; rebuild neighbors/UMAP on the corrected embedding.
9. **Integration checks** — compute **CBF** and **silhouette** (batch/timepoint); visualize pre/post UMAPs.
10. **Clustering & labels** — Leiden clustering; marker DotPlot and signature scores to outline lineages (read-only context for QC).
11. **Biology spot-checks** — ISG score overlays; composition by timepoint/organ; optional within-lineage DE (contextual).
12. **Reference comparison (optional)** — CellTypist mapping with human epithelial model preference; confusion matrix + ARI/NMI + per-lineage precision/recall.
13. **Robustness (compact)** — small grid over neighbors/dimensions to confirm stable integration behavior.
14. **Outputs & manifest** — write figures and tables to `results/`, checkpoints to `data/processed/`, and a session/manifest file to `results/`.

Figures to review

- **QC violins by sample** — distributions after filtering.
- **UMAP pre vs post** — sample separation before; mixed after, with organ/timepoint preserved.
- **CBF & silhouette** — quantitative summary of integration.
- **Marker DotPlot & lineage UMAP** — coherence of epithelial identities.
- **ISG overlays/boxplots** — infection-response signal.
- **(Optional) CellTypist heatmap & UMAP** — agreement with an external reference.

Outputs

- **Figures:** `results/figures/` (UMAPs, DotPlot, compositions, ISG, CBF/silhouette, robustness, optional CellTypist).
- **Tables:** `results/metrics/` (QC thresholds and counts, doublet rates, ambient heuristic, CBF/silhouette CSVs, lineage scores/labels, ISG and DE tables, optional CellTypist metrics).
- **Checkpoints:** `data/processed/` (`.h5ad` objects at key stages: pre-QC → post-QC → integrated).
- **Manifest:** `results/session_and_artifacts.txt` (package versions, AnnData keys, and list of outputs).

Notes

- **Goal of correction:** remove technical sample effects without erasing biological differences; metrics and biology checks are used together to confirm this.
- **Reference mapping (optional):** prefers human intestinal/epithelial models and requires sufficient gene overlap before use; results are summarized with compact metrics and a confusion matrix.