Oncocyrix Multicohort

Oncocyrix Multicohort is a production-grade, Scanpy-based single-cell RNA-seq analysis pipeline designed for **10x Genomics data**, supporting **single-sample and multi-cohort studies**, with optional **CAR-T–aware analysis**, **cell-type annotation**, **pseudobulk DESeq2**, and **multi-database pathway enrichment**.

The pipeline is built for **research reproducibility**, **structured outputs**, and **large-scale cohort integration** (pre/post, multi-patient, multi-condition designs).

Key Features

- ✅**10x Genomics loader (raw mtx/tsv/gz)**
- ✅**Single & multi-cohort modes** (per-sample + combined analysis)
- ✅**Robust QC & filtering** (mitochondrial %, gene counts)
- ✅**Batch correction & integration** (BBKNN)
- ✅**Dimensionality reduction** (PCA, UMAP, optional t-SNE)
- ✅**Clustering & trajectory inference** (Leiden, DPT)
- ✅**Cell-type annotation**
- Metadata-driven (if provided)
- ML-based via **CellTypist** (optional)
- ✅**Cell-type & cluster-specific marker discovery**
- ✅**Group-wise DE analysis** (single-cell level)
- ✅**Pseudobulk aggregation + DESeq2 (via rpy2)**
- ✅**Pathway enrichment** (GO BP/MF/CC, KEGG, Reactome, WikiPathways)
- ✅**Semantic pathway deduplication** (MiniLM + FAISS, optional)
- ✅**Advanced CAR-T state modeling**
- TStemCM, TPEX, TEX, Effector, Proliferation, Exhaustion
- Patient-level pre/post summaries

Installation

Core installation

```
pip install .
```

Full installation (recommended)

Includes CellTypist, BBKNN, GSEApy, DESeq2 glue, and semantic pathway deduplication:

```
pip install ".[all]"
```

Note: DESeq2 requires **R** and the following R packages installed:

```
install.packages(c("DESeq2", "ggplot2", "pheatmap"))
```

Package Structure

```
oncocyrix-multicohort/
├── pyproject.toml
├── README.md
└── oncocyrix_multicohort/
    ├── __init__.py
    └── pipeline.py
```

The CLI entry point is:

```
oncocyrix-multicohort → oncocyrix_multicohort.pipeline:main
```

Usage

Single-sample mode

```
oncocyrix-multicohort
  --mode single
  --single-10x-dir /path/to/10x/sample
  --out-name SC_ANALYSIS_RESULTS
```

Multi-cohort mode (recommended)

```
GSE208653_RAW/
├── metadata.xlsx
├── Pre/
│   ├── Sample1/
│   └── Sample2/
└── Post/
    ├── Sample3/
    └── Sample4/
```

```
oncocyrix-multicohort
  --mode multi
  --multi-base-dir /path/to/GSE208653_RAW
  --out-name SC_ANALYSIS_RESULTS
```

Disable pathway clustering (optional)

```
oncocyrix-multicohort --no-pathway-clustering
```

Metadata File (Multi-mode)

`metadata.xlsx` (optional but strongly recommended):

sample	group	patient_id	cart_phase	response
S1	Pre	P01	Pre	NR
S2	Post	P01	Post	R

Supported columns: - `sample` (required) - `group` (e.g. Pre/Post, Case/Control) - `patient_id` - `cart_phase` - `cart_compartment` - `response`

CAR-T Analysis (Optional)

Enable in code:

```
DO_CART_SCORING = True
```

Outputs: - CAR-T signature scores per cell - Refined CAR-T states (`CART_State_v2`) - UMAPs colored by CAR-T state - Patient-level **Pre vs Post gene summaries** - State-specific marker genes

CAR-T states include: - `TStemCM_like` - `TPEX_like` - `TEX_terminal` - `Effector_TEFF` - `Proliferating_T` - `Terminal_diff`

Output Structure (Combined Analysis)

```
SC_ANALYSIS_RESULTS/
├── 00_analysis_summary/
```

```
├── 01_qc_and_filtering/
├── 02_highly_variable_genes/
├── 03_dimensionality_reduction_and_embeddings/
├── 04_clustering_and_cell_states/
├── 05_celltype_analysis/
├── 05_CART_analysis/
├── 06_groupwise_deg/
├── 07_pathway_enrichment/
├── 08_pseudobulk/
├── 09_reference_summary/
└── pipeline.log
```

Logging

- Console + file logging
- Full R console capture (when using rpy2)
- Detailed pathway deduplication logs

Requirements

- Python $\geq$ 3.9
- R $\geq$ 4.0 (for DESeq2 workflows)
- Recommended RAM $\geq$ 32 GB for large datasets

Citation

If you use this pipeline in your research, please cite:

> Malik S. *Oncocyrix Multicohort: A CAR-T-aware single-cell RNA-seq analysis framework.*

Author

Sheryar Malik
Bioinformatics Scientist

License