Reproducible scRNA-seq Workflow (PBMC)
Technical Summary

Simo Inkala

February 2026

1 Project Objective

This project implements a fully containerized, reproducible single-cell RNA-seq analysis pipeline using Snakemake and Docker. The workflow is demonstrated on publicly available 10x Genomics PBMC datasets (Donors 1–4).

Primary goal: demonstrate production-grade workflow engineering rather than novel biological discovery.

Key design principles:

- Deterministic execution (pinned container + pinned statistical environment)
- Version-locked dependencies (digest-pinned Docker image; `renv.lock`)
- Restart-safe modular execution (`*.done` sentinels; atomic completion markers)
- Donor-aware statistical modeling (pseudobulk DESeq2 with explicit TOST equivalence testing)
- Cross-donor validation (consensus co-expression networks)

2 Architecture

Pipeline layers:

1. Python CLI wrapper (section-based execution)
2. Snakemake DAG (authoritative workflow logic)
3. Docker container (environment isolation)
4. `renv`-locked R environment (statistical layer)

Execution flow:

```
FASTQ → QC → (optional trimming) → STARsolo →
Seurat QC/Clustering → Pseudobulk DE →
Equivalence testing (TOST) → Pathway enrichment →
Consensus co-expression networks
```

Each stage produces atomic `*.done` sentinel files to ensure restart safety and auditability.

3 Reproducibility Strategy

Containerization

- Base image: `rocker/r-ver:4.3.3`
- Digest-pinned release image published to GHCR
- All execution occurs inside Docker (host-independent runtime)

R Environment Control

- `renv.lock` version-locked
- CRAN snapshot pinned (2024-01-01)
- Bioconductor 3.18 pinned
- Seurat / SeuratObject vendored locally (installed from local tarballs)

This eliminates upstream dependency drift and reduces long-term reproducibility risk.

4 Upstream Processing

Data

10x Genomics GEM-X 3′ v4 PBMC libraries (Donors 1–4).

Quality Control

FastQC + MultiQC.

Findings:

- High base quality across donors
- No pervasive adapter contamination
- Trimming unnecessary (optional branch retained for pathological runs)

Alignment

Aligner: STARsolo v2.7.11b
Reference: GRCh38 + GENCODE v45
Chemistry: 16 bp cell barcode + 12 bp UMI

Alignment summary:
Metrics are consistent with high-quality 10x 3′ scRNA-seq data.

Donor	Unique (%)	Multi (%)
Donor1	71.9	20.2
Donor2	72.4	19.3
Donor3	72.4	19.8
Donor4	68.8	21.7

5 Downstream Analysis

Seurat Processing

- `LogNormalize` (scale factor 10,000)
- 2000 HVGs (VST)
- PCA (30 PCs)
- Louvain clustering (resolution 0.3)
- Marker-based annotation via module scores

Major PBMC lineages are recovered consistently across donors.

Pseudobulk Differential Expression (DESeq2)

Counts are aggregated per donor × immune group (donor treated as the experimental unit).

Model:

$$\sim \texttt{donor} + \texttt{group}$$

Tool: DESeq2 v1.42.1

Strict marker definition:

$$\texttt{padj} < 10^{-10}, \quad |\log_2 \mathrm{FC}| > 3$$

Contrasts:

- T-like vs B-like
- T-like vs Mono-like
- B-like vs Mono-like

Equivalence Testing (TOST)

Margin:

$$\delta = 0.75$$

Conserved genes:

$$\texttt{padj}_{equiv} < 0.05 \quad \text{and} \quad |\log_2 \mathrm{FC}| < \delta$$

This separates true small-effect similarity from statistical non-significance, yielding a contrast-specific marker set alongside a shared “core” program.

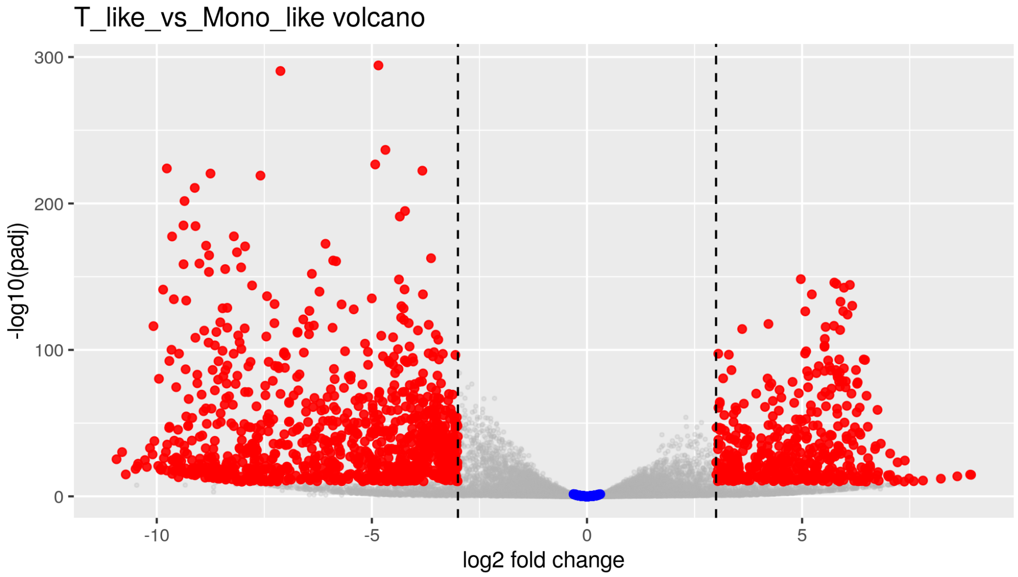

Figure 1: Representative pseudobulk DE contrast (T-like vs Mono-like). Strict markers occupy the distribution tails while near-zero effects concentrate around the origin.

6 Pathway Enrichment

- GSEA (`fgsea`)
- ORA (`clusterProfiler`)
- MSigDB Hallmark + C7

Markers are enriched for inflammatory and immune activation programs (e.g. TNFα/NFκB and interferon signaling), while conserved genes emphasize shared metabolic and proliferative programs (e.g. oxidative phosphorylation, MYC targets, cell cycle).

7 Consensus Co-expression Networks

Constructed separately for:

- CD4 T cells
- B cells
- CD14 monocytes

Method:

- Metacell aggregation (size = 20)
- Spearman correlation across metacells
- Top-k sparsification ($k = 50$) with $|\rho| \geq 0.25$ (positive edges)
- Consensus retention if edge observed in ≥ 2 donors

Leiden clustering is performed on the consensus graph.

8 What This Demonstrates

- Production-grade Snakemake DAG design with restart-safe sentinel outputs
- Containerized statistical reproducibility (`renv` + digest-pinned Docker)
- Dependency pinning and vendoring discipline (Seurat sources)
- Donor-aware modeling (pseudobulk DESeq2) rather than naive per-cell DE
- Explicit equivalence testing (TOST) for principled "no meaningful difference" claims
- Cross-donor consensus network strategy for reproducible co-expression structure

9 Conclusion

This workflow provides a complete, auditable, and reproducible scRNA-seq analysis system from raw FASTQs to donor-aware inference and consensus network construction.

The primary contribution is engineering rigor: reproducibility, modular execution, statistical transparency, and cross-donor validation are treated as first-class design constraints.

Full technical documentation and extended results are available in the accompanying report and repository.

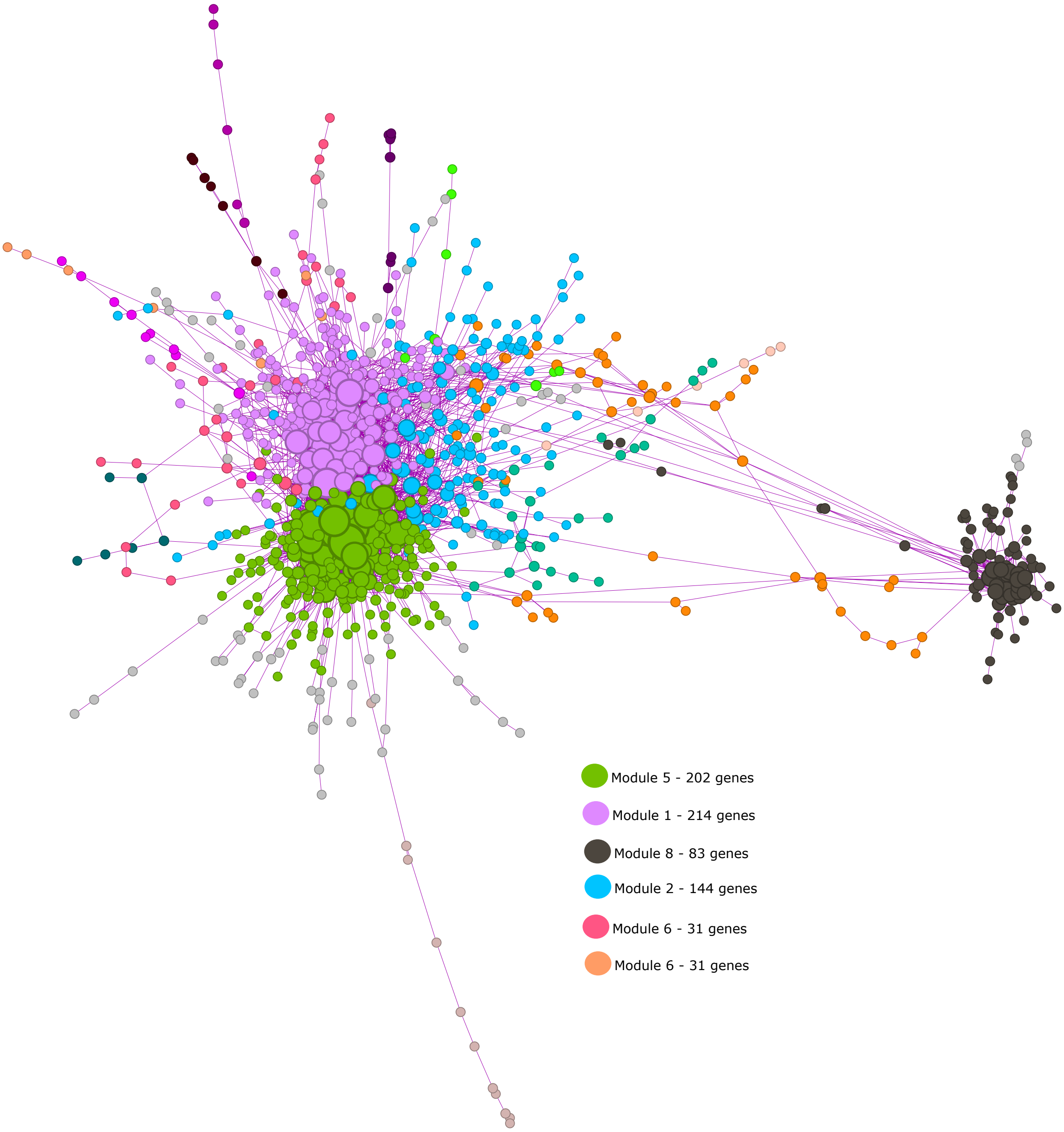

Figure 2: Example consensus network (CD4 T cells) colored by Leiden module. Cross-donor support filtering yields stable modules with clear compartmentalization of inflammatory, interferon, and proliferative programs.