ClusterProfiler-LLM: NSCLC Case Study

Automated Interpretation of Nature Genetics (2025) Data

2026-02-06

Table of contents

1 Introduction

This document demonstrates the capability of `clusterProfiler-LLM` to reproduce and interpret key findings from a high-complexity single-cell study. We analyze the dataset from *Nature Genetics* (2025): "Multi-omic profiling highlights factors associated with resistance to immuno-chemotherapy in non-small-cell lung cancer".

We compare the automated output of our multi-agent framework against the manual conclusions drawn by the original authors, specifically focusing on:

1. **Mechanism Interpretation**: Tumor cell states (e.g., NRF2-mediated stress response).
2. **Cell Type Annotation**: Identification of specific subsets (e.g., SPP1+ Macrophages).
3. **Phenotypic Characterization**: Immune states (e.g., TIGIT/CTLA4-driven exhaustion).

2 Materials and Methods

2.1 Data Source

- **Study**: Yan et al., *Nature Genetics*, 2025.
- **Context**: NSCLC treated with neoadjuvant ICB + chemotherapy.
- **Key Original Findings**:
  - **SPP1+ Macrophages**: Interact with COL11A1+ CAFs to form a physical barrier.
  - **Tumor Cells**: “Basal/Stress” state characterized by NRF2 pathway activation.
  - **T Cells**: Distinct exhaustion trajectories vs. activation.

2.2 Setup

```r
library(clusterProfiler)
library(dplyr)
```

This Quarto document renders from pre-computed objects (RDS) to avoid re-running LLM calls during build. To reproduce the results end-to-end, you will need the development version of `clusterProfiler` that provides `interpret()`, `interpret_agent()`, and `interpret_hierarchical()`, and an LLM backend configured via `fanyi`.

```r
library(fanyi)

api_key <- Sys.getenv("DEEPSEEK_API_KEY")
set_translate_option(source = "dsk", key = api_key)
```

3 Analysis Workflow

3.1 1. Data Preparation

We load the pre-processed marker genes and metadata.

The case study is distributed with the following inputs in the same directory as this QMD:

File	What it contains	Used for
`scobj.markers.rds`	FindAllMarkers output (per-cluster marker genes)	Selecting top markers for enrichment
`md.rds`	Cell-level metadata from the Seurat object	Building context and priors
`x.rds`	Cell-marker enrichment (`compareCluster(..., fun=enricher)`)	LLM interpretation input

File	What it contains	Used for
anno.rds, pheno.rds, deep.rds, hier.rds	Pre-computed LLM outputs	Rendering results without live API calls

The md.rds file is a cell-level metadata table (i.e., scobj@meta.data) saved separately so this case study does not need to bundle the full Seurat object. It is expected to contain, at minimum:

- Study/clinical fields used to summarize dataset context: PathType, Timepoint, PathRes, Drug, Group
- Cluster ID used to aggregate priors: seurat_clusters
- A coarse lineage/cell class label used as prior knowledge: all_cluster_annotation

```
scobj.markers <- readRDS("./scobj.markers.rds")

# Filter for top 10 markers per cluster
scobj.markers %>%
  group_by(cluster) %>%
  dplyr::filter(avg_log2FC > 1) %>%
  slice_head(n = 10) %>%
  ungroup() -> top10

md <- readRDS("./md.rds")
```

3.2 2. Context Construction

We construct a prompt context based on the study design.

```
majority <- function(x) names(sort(table(x), decreasing = TRUE))[1]

ctx <- paste0(
  "Study setting: NSCLC single-cell RNA-seq in the context of neoadjuvant ICB + chemotherapy. '
  "This dataset includes paired/longitudinal sampling (pre- vs post-treatment) and response str
  "Current object summary: ",
  "PathType=", majority(md$PathType), "; ",
  "Timepoint=", majority(md$Timepoint), "; ",
  "PathologicResponse=", majority(md$PathRes), " (e.g., NMPR/MPR/pCR); ",
  "Drug=", majority(md$Drug), "; ",
  "Group=", majority(md$Group), ". ",
  "Goal: annotate clusters/cell states and interpret marker-enrichment results in this context.
)
```

We also construct two optional inputs used by the interpretation functions:

1. A prior vector (prior_vec) derived from the dataset's coarse lineage annotation (all_cluster_annotation).
2. A per-gene fold-change lookup (gene_fold_change) derived from marker statistics.

```r
prior_tbl <- md %>%
  mutate(cluster = .data[["seurat_clusters"]]) %>%
  group_by(cluster) %>%
  summarise(prior = majority(all_cluster_annotation), .groups = "drop")
prior_vec <- setNames(prior_tbl$prior, as.character(prior_tbl$cluster))

fc_col <- if ("avg_log2FC" %in% names(top10)) "avg_log2FC" else "avg_logFC"
gene_fc <- top10 %>%
  group_by(gene) %>%
  summarise(fc = max(.data[[fc_col]], na.rm = TRUE), .groups = "drop")
gene_fold_change <- setNames(gene_fc$fc, gene_fc$gene)
```

3.3 3. Standard Enrichment Analysis

```r
x <- readRDS("./x.rds")
```

The enrichment objects were generated from `top10` using `compareCluster`. Cell-marker enrichment used an external TERM2GENE table (`Cell_marker_Human.xlsx`), which is not bundled in this directory.

```r
cm <- readxl::read_xlsx("./Cell_marker_Human.xlsx")

x <- compareCluster(
  gene ~ cluster,
  data = top10,
  fun = enricher,
  TERM2GENE = cm[, c("cell_name", "marker")]
)
saveRDS(x, "./x.rds")
```

4 Results: LLM-Driven Interpretation

4.1 Task 1: Cell Type Annotation (Hypothesis Verification)

We use `interpret(task = "celltype")` with prior knowledge injection.

```r
library(fanyi)

api_key <- Sys.getenv("DEEPSEEK_API_KEY")
set_translate_option(source = "dsk", key = api_key)

anno <- interpret(
  x = x,
  context = ctx,
  task = "celltype",
  prior = prior_vec,
  n_pathways = 20,
  add_ppi = FALSE,
  gene_fold_change = gene_fold_change,
  model = "deepseek-chat",
  api_key = api_key
)
saveRDS(anno, "./anno.rds")
```

Result:

Comparison with *Nature Genetics* (2025):

Biological Feature	clusterProfiler-LLM Interpretation	Nature Genetics (Original Findings)	Key Match?	Significance
Macrophage Identity	Identified **SPI1+ M2-like Macrophages** with immunosuppressive signatures (*C1QA/B/C*, *CD163*).	Described **SPP1+ TAMs** as a major immunosuppressive population interacting with CAFs.	**Yes** (M2/TAM overlap)	Major Resistance Mechanism
Tumor Cell State	Detected **Basal-like Progenitor** state with *GPX2* and *AKR1C1* (Oxidative Stress).	Defined **"Basal/Stress"** tumor cell state characterized by **NRF2** activation.	**Direct** (Stress/Basal)	Poor Prognosis Driver
Stromal Niche	Annotated **Activated Myofibroblasts** (*ACTA2*, *COL1A2*) and Vascular Progenitors.	Detailed **COL11A1+ CAFs** forming a barrier at tumor boundaries.	**Yes** (CAF Activation)	Immune Exclusion
T Cell States	Found **IL7R+ Memory T cells** and **Exhausted** subsets.	Highlighted dysfunctional CD8+ T cells and potential for revival.	**Yes**	Therapy Response Determinant

4.2 Task 2: Phenotypic Characterization

We use `interpret(task = "phenotype")` to define the functional state of each cluster.

```r
library(fanyi)

api_key <- Sys.getenv("DEEPSEEK_API_KEY")
set_translate_option(source = "dsk", key = api_key)

pheno <- interpret(
  x = x,
  context = ctx,
  task = "phenotype",
  n_pathways = 30,
  add_ppi = FALSE,
  gene_fold_change = gene_fold_change,
  model = "deepseek-chat",
  api_key = api_key
)
saveRDS(pheno, "./pheno.rds")
```

Result:

Comparison with *Nature Genetics* (2025):

Biological Feature	clusterProfiler-LLM Interpretation	Nature Genetics (Original Findings)	Key Match?	Significance
Tumor Cell Plasticity	Identified **Basal-like Progenitor** state expressing *AKR1C1*, *GPX2* (Detoxification/Stress).	Defined **"Basal/Stress"** state driven by NRF2 (*NFE2L2*) and oxidative stress response.	**Direct** (AKR1C1/Stress)	Chemoresistance Mechanism
Fibroblast Activation	Characterized **Activated Myofibroblasts** with high *COL1A1*, *ACTA2*, *TGFB1*.	Described **COL11A1+ CAFs** forming immune-excluded niches.	**Yes** (TGF-( ) CAFs)	Immune Exclusion Barrier

Biological Feature	clusterProfiler-LLM Interpretation	Nature Genetics (Original Findings)	Key Match?	Significance
Myeloid Polarization	Detected **Immunosuppressive Macrophages** (*C1QA*, *APOE*) with M2-like features.	Highlighted **SPP1+ TAMs** as the dominant immunosuppressive myeloid population.	**Yes** (M2/TAMs)	T-cell Suppression
B Cell Function	Noted **Germinal Center B cells** (*BCL6*, *AICDA*) indicating humoral immunity.	Discussed **TLS (Tertiary Lymphoid Structure)** presence and B cell maturity.	**Consistent**	Anti-tumor Immunity Potential

4.3 Task 3: Mechanism Interpretation (Multi-Agent Deep Mode)

We use `interpret_agent()` to reconstruct causal networks, integrating PPI and LogFC data.

```r
library(fanyi)

api_key <- Sys.getenv("DEEPSEEK_API_KEY")
set_translate_option(source = "dsk", key = api_key)

deep <- interpret_agent(
  x = x,
  context = ctx,
  n_pathways = 50,
  add_ppi = TRUE,
  gene_fold_change = gene_fold_change,
  model = "deepseek-chat",
  api_key = api_key
)
saveRDS(deep, "./deep.rds")
```

Result:

Comparison with *Nature Genetics* (2025):

Biological Feature	clusterProfiler-LLM Interpretation	Nature Genetics (Original Findings)	Key Match?	Significance
T Cell Exhaustion	**TOX** identified as a key regulator of **Exhausted CD8+ T cells** (*PDCD1*, *HAVCR2*).	**Dysfunctional CD8+ T cells** are a major feature of non-responders; exhaustion is a key barrier.	**Direct** (TOX/Exhaustion)	Checkpoint Blockade Target
Myeloid Regulation	**SPI1 (PU.1)** inferred as the master regulator for **M2-like TAMs** (*C1QA*, *MRC1*).	**SPP1+ TAMs** recruit regulatory T cells and suppress adaptive immunity.	**Yes** (Myeloid Driver)	Immunosuppressive Hub
B Cell Identity	**PAX5** and **BCL6** regulatory network defines **Follicular/GC B cells**.	Presence of mature B cells in **TLS** correlates with better prognosis.	**Consistent**	Prognostic Marker
Tumor Proliferation	**E2F** targets and **MYC** signaling active in **Cycling Tumor Cells**.	High proliferation rates in specific tumor subclones drive progression.	**Yes**	Tumor Aggressiveness

4.4 Task 4: Hierarchical Interpretation

Refining annotations from Major lineage to Minor states using `interpret_hierarchical()`.

```r
library(fanyi)

api_key <- Sys.getenv("DEEPSEEK_API_KEY")
set_translate_option(source = "dsk", key = api_key)

map_tbl <- md %>%
  mutate(minor = as.character(minor), major = as.character(major)) %>%
  count(minor, major, name = "n") %>%
  group_by(minor) %>%
  slice_max(n, n = 1, with_ties = FALSE) %>%
  ungroup()
mapping <- setNames(map_tbl$major, map_tbl$minor)

top10_minor <- readRDS("./top10_minor.rds")
top10_major <- readRDS("./top10_major.rds")
```

```r
cm <- readxl::read_xlsx("./Cell_marker_Human.xlsx")

x_minor <- compareCluster(
  gene ~ cluster,
  data = top10_minor,
  fun = enricher,
  TERM2GENE = cm[, c("cell_name", "marker")]
)
saveRDS(x_minor, "./x_minor.rds")

x_major <- compareCluster(
  gene ~ cluster,
  data = top10_major,
  fun = enricher,
  TERM2GENE = cm[, c("cell_name", "marker")]
)
saveRDS(x_major, "./x_major.rds")

hier <- interpret_hierarchical(
  x_minor = x_minor,
  x_major = x_major,
  mapping = mapping,
  model = "deepseek-chat",
  api_key = api_key,
  task = "cell_type"
)
saveRDS(hier, "./hier.rds")
```

Result:

Comparison with *Nature Genetics* (2025):

Biological Feature	clusterProfiler-LLM Interpretation	Nature Genetics (Original Findings)	Key Match?	Significance
CD8+ T Cell Subsets	Distinctly separated **Naïve**, **Memory** (*IL7R*), and **Exhausted** (*TOX*, *LAG3*) T cells.	Emphasized the spectrum from **Pre-dysfunctional** to **Dysfunctional** CD8+ T cells.	**Precise**	Therapy Response Continuum

Biological Feature	clusterProfiler-LLM Interpretation	Nature Genetics (Original Findings)	Key Match?	Significance
Macrophage Heterogeneity	Hierarchically resolved **Alveolar Macrophages** vs. **Tumor-Associated Macrophages** (*C1QA*).	Distinguished resident macrophages from tumor-infiltrating **SPP1+ TAMs**.	**Yes**	Origin Matters (Tissue vs. Tumor)
Tumor Heterogeneity	Sub-classified tumor cells into **Cycling** (*MKI67*) and **Stress/Basal** (*AKR1C1*) states.	Mapped tumor macro-clusters to distinct **Cellular States** (Cycling, Stress, Interferon-high).	**Direct**	Intratumoral Heterogeneity
Endothelial States	Identified **Tip cells** and **Stalk cells** indicative of angiogenesis.	Noted **PLVAP+ Endothelial cells** associated with tumor vascularization.	**Consistent**	Angiogenesis Targets

5 Conclusion

The `clusterProfiler-LLM` framework successfully reproduced the key biological insights of the 2025 *Nature Genetics* study without manual curation. By automating the interpretation of mechanism, cell identity, and phenotype, it accelerates the transition from data to knowledge.